BOLT BERANEK AND NEWMAN INC

CONSULTING · DEVELOPMENT · RESEARCH

AFCRL-68-0446

30 August 1968

STUDY OF ACOUSTIC PROPERTIES OF SPEECH SOUNDS

Kenneth N. Stevens
Gary M. Klatt

Scientific Report No. 8
Contract No. F19628-68-C-0125
Project No. 8668
Contract Monitor: Hans Zschirnt, Data Sciences Laboratory

Prepared for:

AIR FORCE CAMBRIDGE RESEARCH LABORATORIES
Office of Aerospace Research
United States Air Force
Bedford, Massachusetts 01730

This research was sponsored by the Advanced Research Projects Agency under ARPA Order No. 627.

AFCRL-68-0446 30 August 1968

STUDY OF ACOUSTIC PROPERTIES OF SPEECH SOUNDS

Kenneth N. Stevens
Mary M. Klatt

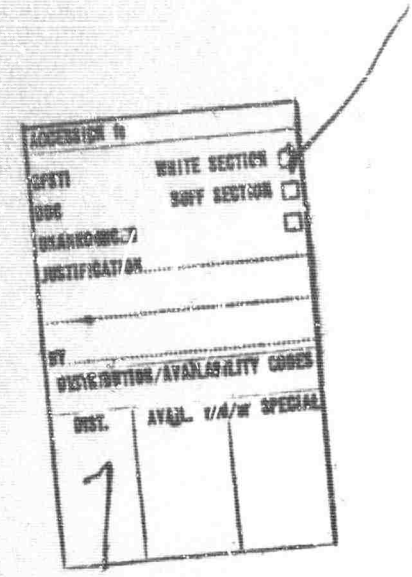
BOLT BERANEK AND NEWMAN INC
50 Moulton Street
Cambridge, Massachusetts 02138

Scientific Report No. 8
Contract No. F19628-68-C-0125
Project No. 8668
Contract Monitor: Hans Zschirnt, Data Sciences Laboratory

Prepared for:

AIR FORCE CAMBRIDGE RESEARCH LABORATORIES
Office of Aerospace Research
United States Air Force
Bedford, Massachusetts 01730

This research was sponsored by the Advanced Research Projects Agency under ARPA Order No. 627.



ABSTRACT

The spectral and temporal characteristics of American English vowel and consonant sounds in a variety of phonetic contexts are examined and compared with data reported in the literature. Spectrograms and sampled spectra (obtained from an analog filter bank connected to a digital computer) were assembled for a number of monosyllabic and bisyllabic utterances generated by three talkers, and a variety of measurements were made from these displays. The characteristics examined include durations of vowels, durations of various phases of consonants in prestressed and poststressed positions and in clusters, spectra of vowels and diphthongs and their variation with time, spectra of consonants during constricted intervals, and time-variation of spectra during the release of consonants. The aim of the study is not to present an exhaustive acoustic-phonetic description of American English speech sounds but rather to indicate the kinds of acoustic properties that need to be utilized in schemes for machine recognition of speech.

TABLE OF CONTENTS

		page
Abstract		iii
List of Figures		v
List of Tables		viii
1.	Introduction	1
2.	Procedures for Processing the Data	5
3.	Description of Speech Material	10
4.	Stressed Vowels	15
5.	Consonants in Prestressed Position: Single Consonants	35
	5.1 Closure interval: stop and nasal consonants	37
	5.2 Constricted interval: fricative consonants	43
	5.3 Constricted interval: liquids and glides (sonorant, nonnasal consonants)	49
	5.4 Release and transitions: stop and nasal consonants	53
	5.5 Release and transitions: liquids and glides	62
	5.6 Summary of characteristics of consonants in prestressed position	62
6.	Consonant Clusters in Prestressed Position	66
7.	Unstressed Vowels and Vowels with Secondary Stress	69
	7.1 Duration and fundamental frequency	70
	7.2 Spectra of vowels with secondary stress	71
	7.3 Spectra of reduced vowels	73
8.	Consonants in Poststressed Positions	75
9.	Concluding Remarks	81
References		83

LIST OF FIGURES

page

Figure 1. Measured frequency response of filter bank. Ratio of dc output to ac input in dB 7

2. Impulse response of low-pass filter used to smooth rectified outputs of band-pass filters . 8

3. Spectrogram of the word *raucous* (speaker KS) and Printout of 19-channel filter bank for the word *raucous* 9

4. Spectrograms of five vowels in the environment /b-b/, generated by speaker KS 17

5. Spectra of 15 stressed vowels in the environment /b-b/ obtained from 19-channel filter bank 22

6. Spectra of three stressed vowels (in the environment /b-b/) are compared for three speakers. Data obtained from 19-channel filter bank 30

7. The upper graph shows the range of spectra for the stressed vowel /ɪ/ for seven different consonantal environments in nonsense syllables and in bisyllabic words. The lower two graphs are examples of spectra of the same vowel when it is modified appreciably by the final consonant. Speaker KS 32

8. Spectrograms illustrating properties of stop and nasal consonants. Speaker KS 38

9. Spectra during closure interval for voiced stop consonants obtained from 19-channel filter bank 40

10. Spectra during closure for nasal consonants in prestressed position 41

11. Spectrograms illustrating properties of voiceless fricative consonants. Speaker KS 44

Figure 12. Spectra during constricted interval for voiceless fricative consonants in prestressed position. Speaker KS 45

13. Spectrograms illustrating properties of voiced fricative consonants. Speaker KS 47

14. Spectra during constricted interval for voiced fricative consonants in prestressed position. Speaker KS 48

15. Spectrograms illustrating properties of liquids and glides. Speaker KS 50

16. Spectra sampled during middle of constricted interval for liquids and glides in the prestressed position. Speaker KS 52

17. Graph of amplitude of channel 5 *vs* time for the utterances /ə'mɑ/ and /ə'wɑ/ 55

18. Output of channel 2 of spectrum analyzer as a function of time during the release of three cognate pairs of voiced and voiceless stop consonants. Speaker KS 57

19. Outputs of filters 5, 8, and 16 during time interval immediately preceding and following the release for the syllables indicated. Speaker KS 59

20. Spectra (obtained from the 19-channel filter bank) sampled at 60-msec intervals following release of the stop consonants in the syllables /pɑ/, /tɑ/, and /kɑ/. Speaker KS 61

21. Outputs of several filters (as indicated) for liquid and glide consonants in the environment /ə'Cɑ/. Speaker KS 63

22. Examples of spectrograms of consonant clusters in prestressed position. Speaker KS 67

23. Spectra of vowels with secondary stress 72

24. Spectra of unstressed vowels that are reduced. Speaker KS 74

Figure 25. Spectrograms illustrating acoustic characteristics of consonants in poststressed position preceding reduced vowels. Speaker KS 76

26. Spectra of some voiced consonants during the closure interval. The consonants are in poststressed position preceding reduced vowels. Shown for comparison are spectra of the same consonants in the prestressed environment /ə'Cɑ/: Speaker KS 78

LIST OF TABLES

		page
I.	List of filter center frequencies and bandwidths ...	6
II.	List of utterances used in phonetic study	12
III.	Features of the vowels of American English	15
IV.	Durations of stressed vowels in the environment /bVb/. Averages for three talkers generating one utterance for each vowel	18
V.	Measured vowel durations (in milliseconds) from vowels and vowel-consonant combinations occurring in bisyllabic words spoken in isolation	20
VI.	Classifications of consonants in terms of distinctive features	36
VII.	Durations of closure intervals for stop and nasal consonants preceding stressed vowels. Averages over three vowel environments /i ɑ u/ and over three talkers	42
VIII.	Durations of constricted intervals for fricative consonants preceding stressed vowels. Averages over three vowel environments /i ɑ u/ and over three speakers	49
IX.	Durations of noise in fricative /s/ and of stop gap or sonorant murmur in following consonant for various consonant clusters in the environment /ə'sC(C)ɑ/. Average values for three speakers	68

1. INTRODUCTION

The purpose of this report is to summarize some data on the acoustic properties of speech sounds of American English. The motivation is to present information that may be useful to the researcher who is interested in machine recognition of speech.

A great deal of information with regard to the acoustic properties of speech sounds has been published in journals and books devoted to acoustics and phonetics. This published material is not, however, directly usable by those engaged in automatic speech recognition for several reasons. A principal reason is that the data reported in the past have often not been presented in sufficiently quantitative form. In many cases, data have not been given in absolute terms, but rather relative values of properties have been reported. Furthermore, numerical data are frequently obtained from spectrograms or other displays in which a human observer must interpret the display in order to make a measurement. In an application such as machine recognition of speech, it is, of course, essential that all analysis be done by machine. It is by no means obvious that a machine can be programmed to perform the same kinds of analysis as a human observer.

For these reasons, our study of the acoustics of speech has included not only an examination of existing phonetics information but also the acquisition and interpretation of some new data. In our analysis of these data, we have attempted to specify acoustic properties in a reasonably quantitative way that is amenable to machine processing.

Our point of view in this study is that each phonetic unit is characterized by a set of underlying attributes or features.

These features have certain well-defined articulatory and acoustic correlates. When a phonetic segment is concatenated with other phonetic segments, the articulatory and acoustic properties may become distorted or modified as a consequence of the phonetic environment. Thus a study of the acoustic characteristics of phonetic segments must include not only an examination of the underlying "undistorted" attributes of the segments but also a consideration of the effects of the various phonetic environments in which the segment can appear.

It is probable that the properties of a phonetic segment are modified least by the environment when the segment occurs in an isolated, stressed, consonant-vowel syllable. Acoustic invariants for a consonant are more likely to be observed when it is the only consonant preceding the stressed vowel in the syllable. It might be hypothesized that a consonant or vowel in such a syllable would have acoustic characteristics that are closest to the "ideal" characteristics. In general, nonsense utterances of the form /ə'CVC/ (C = consonant, V = vowel)* were used in this study to obtain data corresponding to this ideal situation. The final consonant may have some influence on the characteristics of the preceding vowel, but it is possible to select consonants whose effect on the vowel is minimal.

The modifications that occur in a segment as a consequence of its phonetic environment are of several types. A stressed vowel undergoes some change if the syllable in which it occurs is not in isolation or, in general, is not in the final position of an utterance.

---

*The apostrophe indicates that the stress is on the second syllable. The stress pattern in these nonsense utterances is like that in the word *about*.

Such effects are examined in a preliminary way in this study by obtaining data from bisyllabic words which have the stress on the first syllable. Furthermore, stressed vowels are modified by the final consonants in the syllables in which they occur, and consequently it is necessary to examine the characteristics of vowels with many following consonants. Consonants in prestressed position can undergo appreciable modification when they occur in consonant clusters, and utterances of the form /ə'$C_1C_2(C_3)V$/ are used to examine these effects. Likewise, consonant characteristics may be altered when they are in poststressed position. Some utterances to illustrate and, where possible, to quantify these effects are included in the corpus of material in this study.

A goal in any study of the acoustic properties of speech sounds is to find a description of a phonetic unit (or, for that matter, of a group of a small number of units, such as a syllable) that is sufficient to permit that unit to be uniquely identified without reference to the context in which it appears. It must be recognized, however, that the nature of human speech precludes the achievement of this ideal goal. It is common for a given phonetic unit to be so distorted in continuous speech that acoustic data in the speech signal in the vicinity of that unit are insufficient to provide an unequivocal identification of the unit. A listener is able to make an interpretation of an utterance in which this unit appears because he is familiar with the rules governing the sequence of phonetic units (or, more precisely, the matrix of phonetic features) that can occur in his language. In cases where a phonetic unit is not sufficiently well defined acoustically, the listener must make use of these rules to infer the presence and the identity of this unit.

Thus a study such as the present one must not be considered as a complete description of a set of phonetic units that will specify algorithms for recognizing these units, but rather must be viewed as leading towards procedures for preprocessing the speech signal, to obtain as much information as possible from the acoustic waveform. In a speech-recognizing device, the results of this preprocessing would then be subjected to further analysis or interpretation by a component in which the lexicon, the phonological rules, and even certain syntactic and semantic rules of the language are stored.

2. PROCEDURES FOR PROCESSING THE DATA

The data presented here were obtained from recordings of a number of monosyllabic and bisyllabic utterances of three talkers – two males (KS and CW) and one female (GC). These utterances were subjected to two kinds of preliminary processing. First, wide-band spectrograms of all the words were made by using a Voiceprint Laboratories sound spectrograph. The so-called logarithmic frequency display was used, covering the frequency range to 7000 Hz. Secondly, all utterances were passed through a specially designed 19-channel filter bank, and the rectified and smoothed outputs of the filter bank were sampled and quantized (on a logarithmic scale) and stored in the memory of a PDP-1 computer. The amplitude quantization is such that each (logarithmic) amplitude step represents 3/4 dB. The numerical values of the sampled spectra so obtained were printed out to permit detailed examination and analysis.

The characteristics of the filter bank have been described elsewhere (Stevens and von Bismarck, 1967). Table I lists the center frequencies and bandwidths of the filters. Up to about 3000 Hz, the filters have bandwidths of 360 Hz and are spaced 180 Hz apart. At higher frequencies, the bandwidths are greater, and the frequency responses of adjacent filters overlap at the 3-dB points. Figure 1 shows the frequency-response curves for the filters, and Fig. 2 shows the impulse response of the low-pass smoothing filter in each channel. From Fig. 2 it can be seen that the low-pass filters average the rectified outputs of the bandpass filters over a time interval of 10–15 msec. This averaging time has, of course, an important influence on the observed characteristics of rapidly changing sounds, such as the onset of stop consonants.

TABLE I. List of filter center frequencies and bandwidths.

Filter No.	Lower Cutoff (Hz)	Higher Cutoff (Hz)	Center Frequency (Hz)	Band-width (Hz)
1	100	440	260	360
2	260	620	440	360
3	440	800	620	360
4	620	980	800	360
5	800	1160	980	360
6	980	1340	1160	360
7	1160	1520	1340	360
8	1340	1700	1520	360
9	1520	1880	1700	360
10	1700	2060	1880	360
11	1880	2240	2060	360
12	2060	2420	2240	360
13	2240	2600	2420	360
14	2420	2780	2600	360
15	2600	2960	2780	360
16	2960	3560	3260	600
17	3560	4400	3980	840
18	4400	5480	4940	1080
19	5480	6560	6020	1080

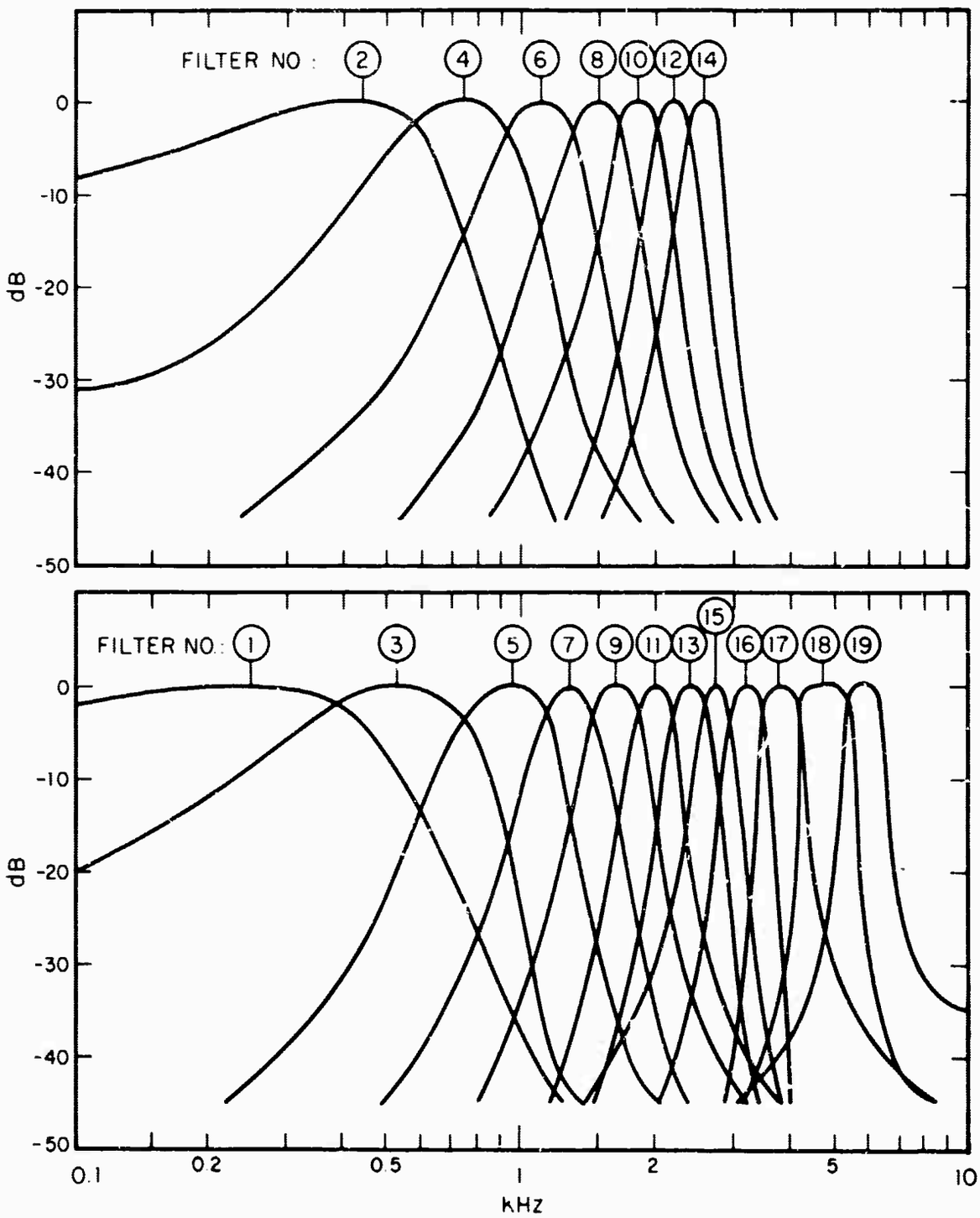


Fig. 1 Measured frequency response of filter bank. Ratio of dc output to ac input in dB.

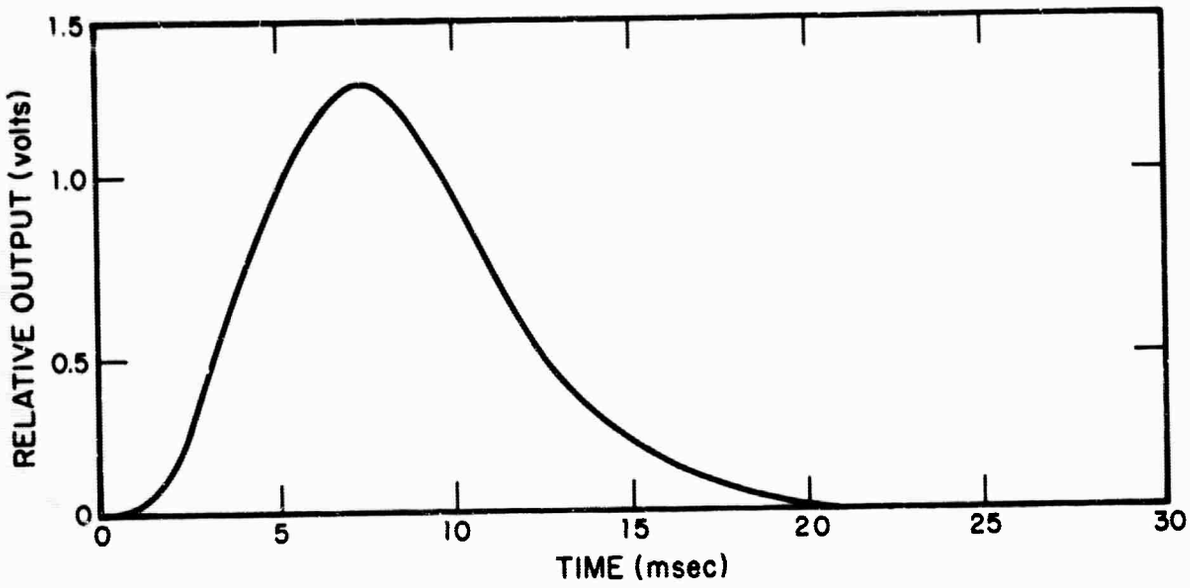


Fig. 2 Impulse response of low-pass filter used to smooth rectified outputs of band-pass filters.

An example of a spectrogram and a printout for one of the utterances is shown in Fig. 3. Major characteristics of the utterance are observable on both displays: the transition from the initial /r/ to the vowel (samples 10–20), the stop gap for the /k/ (samples 37–42), the aspiration following the /k/ release (samples 43–45), and the final /s/ (samples 55–80) can all be identified on the spectrogram and on the printout. The spectrographic display reveals patterns that are interpretable visually by the human observer, whereas the computer printout provides a display that a human can interpret only with difficulty, but which is suitable for machine processing.

word 295 speaker 2 repetition 1

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	total
1	12	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	14
2	13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	13
3	9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9
4	17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	17
5	15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	15
6	11	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	12
7	17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	17
8	17	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	22
9	26	20	9	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	58
10	31	27	12	3	2	0	0	0	0	0	0	0	0	0	0	0	0	0	1	76
11	36	32	19	5	2	1	0	0	0	0	0	0	0	0	0	0	0	0	1	96
12	36	33	20	7	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	99
13	40	36	24	10	3	1	0	0	0	0	0	0	0	0	0	0	0	0	0	114
14	40	38	27	14	9	2	0	0	0	0	0	0	0	0	0	0	0	0	0	130
15	42	40	30	20	18	7	1	0	0	0	0	0	0	0	0	0	0	0	0	158
16	42	42	32	21	20	13	4	2	0	0	0	0	0	0	0	0	0	0	0	176
17	42	43	34	23	27	24	10	6	0	0	0	0	0	0	0	0	0	0	0	209
18	42	44	36	25	28	26	17	13	1	0	0	0	0	0	0	0	0	0	0	232
19	44	47	42	31	32	30	22	19	7	1	0	0	0	0	0	0	0	0	2	277
20	47	50	48	39	34	32	23	20	12	6	2	0	1	0	0	1	2	0	4	321
21	49	55	54	47	39	36	26	19	13	10	9	5	3	0	1	2	0	0	10	378
22	49	57	57	50	42	36	27	22	16	16	14	9	4	0	0	5	1	0	15	422
23	47	56	56	52	46	42	28	23	19	18	18	16	7	0	1	7	3	0	14	453
24	47	56	57	53	48	43	29	23	19	21	21	19	10	0	3	10	4	0	14	477
25	47	56	57	55	51	46	30	22	20	23	24	23	14	3	6	13	3	2	15	510
26	45	55	57	56	52	47	31	21	18	23	24	23	17	5	6	13	3	1	16	513
27	43	54	56	56	53	47	32	22	17	23	24	24	18	7	6	14	3	3	17	519
28	44	53	56	57	55	49	33	23	17	23	23	24	19	9	7	17	7	5	15	536
29	44	53	56	57	56	50	35	24	17	22	23	24	20	11	7	18	9	8	16	550
30	44	52	55	57	56	51	36	25	17	22	23	24	21	14	10	17	11	7	14	556
31	46	52	54	56	56	51	36	25	17	22	23	24	20	11	8	18	8	5	16	548
32	46	52	53	54	55	52	39	28	19	22	22	23	21	19	16	16	8	4	12	561
33	45	51	52	52	52	48	36	26	18	20	19	19	18	13	10	19	8	2	9	517
34	45	50	48	48	45	40	38	29	18	19	17	18	25	21	14	17	11	2	10	523
35	45	48	45	44	45	42	35	26	16	16	12	12	17	13	8	14	10	1	8	459
36	38	44	42	40	42	38	33	28	13	11	8	8	12	8	5	9	5	0	5	389
37	26	40	30	36	36	33	27	20	7	7	5	5	8	3	3	5	3	0	2	302
38	24	35	32	32	32	28	25	19	5	4	3	2	0	1	1	3	2	0	2	256
39	21	31	27	27	28	24	20	13	2	3	2	1	4	0	1	2	1	0	1	208
40	16	26	23	24	24	19	17	9	2	2	1	0	3	0	0	1	1	0	2	170
41	16	23	19	19	20	15	13	6	1	1	1	0	2	0	0	1	0	0	0	137
42	19	20	17	17	18	14	11	11	7	5	2	0	2	0	0	1	2	0	0	146
43	26	25	25	24	23	24	20	32	27	23	14	9	11	4	4	4	10	0	1	312
44	25	27	25	20	17	17	19	25	20	19	8	5	7	1	2	4	10	2	4	253
45	24	26	26	18	14	13	15	24	20	19	8	3	7	1	2	3	7	2	3	234
46	46	42	36	29	21	21	22	29	23	18	14	9	9	3	4	8	9	1	4	348
47	45	46	43	35	24	26	28	33	24	17	14	11	11	6	6	15	9	0	3	396
48	44	48	46	39	28	28	30	34	23	17	14	12	14	9	7	17	9	0	1	418
49	42	47	45	38	25	28	30	33	22	15	12	11	13	9	8	17	9	0	3	407
50	42	47	45	38	25	28	30	32	20	13	11	10	12	9	9	17	9	0	3	400
51	39	45	44	36	24	27	29	30	18	10	9	9	11	9	10	17	10	0	2	379
52	39	44	42	35	23	25	27	27	15	7	7	6	9	6	8	16	9	0	2	345
53	38	42	40	32	20	23	24	24	10	5	5	5	6	3	7	15	9	0	0	308
54	32	36	34	27	15	17	19	19	6	3	3	2	4	0	4	9	7	0	4	241
55	32	32	31	23	10	11	16	15	4	2	2	1	4	0	5	10	16	12	19	245
56	29	28	26	19	8	8	12	11	3	2	3	2	4	0	5	17	23	24	32	256
57	14	24	22	15	6	4	8	7	3	2	3	1	4	0	3	18	26	27	34	221
58	17	19	17	8	4	2	5	4	2	1	3	0	2	0	2	15	27	31	40	199
59	20	17	12	6	2	1	3	3	2	1	2	0	2	0	2	18	31	32	42	196
60	17	16	8	4	2	0	2	2	2	1	3	0	2	0	1	19	31	34	42	186
61	20	13	6	3	2	1	2	2	2	1	3	0	3	0	2	20	35	35	41	191
62	15	13	10	10	5	2	3	2	2	1	4	1	5	0	3	23	33	35	43	210
63	12	10	5	3	3	1	2	1	2	1	3	1	3	0	2	24	34	35	42	184
64	16	6	3	2	2	0	1	1	2	1	3	1	3	0	3	26	36	35	42	185
65	12	6	2	2	2	0	1	1	2	1	3	2	4	0	4	26	36	34	46	184
66	15	5	2	2	2	0	1	2	3	2	4	2	5	0	7	32	39	35	43	201
67	17	2	2	2	2	1	2	2	3	2	4	1	6	2	8	35	38	33	42	204
68	20	0	1	1	2	0	1	1	2	1	3	1	5	1	6	31	36	33	43	194
69	10	4	1	1	2	1	1	1	2	1	3	1	5	0	6	34	36	34	44	187
70	12	7	5	5	3	2	2	2	2	1	4	3	7	2	11	34	35	32	43	212
71	20	4	4	3	3	2	2	2	2	1	4	4	7	6	15	34	36	34	40	223
72	13	5	6	7	5	2	2	3	3	2	7	11	13	7	17	31	34	31	36	235
73	11	4	5	6	4	3	2	2	2	1	3	5	8	6	16	29	32	28	35	201
74	8	1	3	3	2	1	1	2	1	1	4	4	7	4	15	26	29	24	32	168
75	17	2	3	3	2	2	2	2	0	0	2	3	5	2	12	23	24	21	29	155
76	10	1	3	2	2	2	2	2	1	1	3	2	5	2	12	23	24	18	26	141
77	18	5	4	3	2	1	2	2	1	1	2	1	5	2	10	21	19	13	23	137
78	22	3	4	3	3	3	3	3	2	0	1	0	4	1	8	17	15	9	18	119
79	16	1	2	2	2	3	3	4	0	0	0	0	1	0	3	16	10	5	14	82
80	14	3	3	2	2	2	3	3	1	0	0	0	0	0	2	10	6	3	10	64
81	13	1	2	2	2	3	3	3	1	0	0	0	0	0	2	8	5	2	7	54
82	18	6	3	2	4	7	6	5	1	0	0	0	0	0	1	6	3	1	5	68

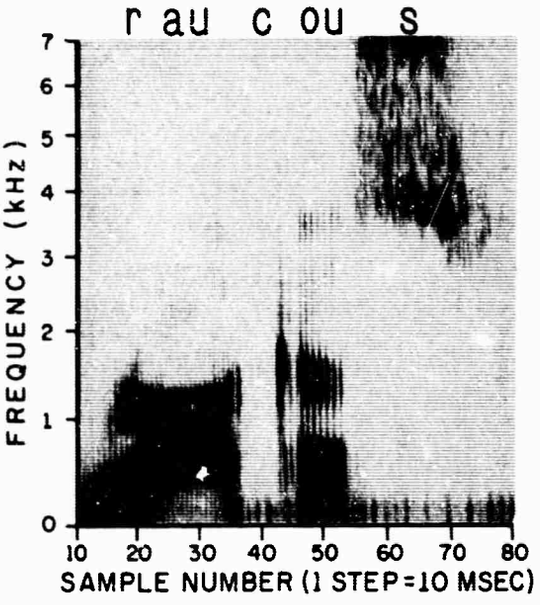


Fig. 3

Above: Spectrogram of the word *raucous* (speaker KS). The samples indicated on the abscissa represent instants of time at which the outputs of the filter bank were sampled to obtain the print-out at the left.

Left: Printout of 19-channel filter bank for the word *raucous*. The sample numbers are designated in the left-hand column, and the "total" at the right represents the sum of all filter outputs. The numbers represent the amplitudes of the filter outputs on a logarithmic scale in steps of approximately 3/4 dB.

3. DESCRIPTION OF SPEECH MATERIAL

The utterances used in this study are listed in Table II.* This particular selection of utterances was made in order to sample a wide variety of speech sounds occurring in various phonetic contexts. Some of the material consists of nonsense syllables or bisyllables, and other utterances are words of English. The series of syllables of the form bVb was included to obtain basic data on all the vowels and diphthongs in a consonantal environment that was considered to have a minimal influence on the vowel. The nonsense utterances of the form /ə'CV(C)/ were used to obtain further data on vowels in different consonantal environments as well as to examine various consonants in prestressed position and in final position. Consonant clusters are represented in utterances of the form /ə'C_1C_2V/ or /ə'$C_1C_2C_3$V/. The bisyllabic English words were used to provide examples of vowels and consonants in other phonetic environments, particularly unstressed vowels and consonants in poststressed position. Some real monosyllabic and bisyllabic words were included to provide examples of consonant clusters in poststressed position.

The material in this report is organized into four parts, each part being concerned with a particular class of speech sounds:

---

*The phonetic symbols used in Table II and throughout the text are (with some minor modifications) the symbols of the International Phonetic Association. Examples of words containing the nonobvious phonetic symbols are the following: /i/ (beet), /ɪ/ (bit), /e/ (bait), /ɛ/ (bet), /æ/ (bat), /ɑ/ (cot), /ʌ/ (cut), /ɔ/ (bought), /o/ (boat), /ʊ/ (foot), /u/ (boot), /ɑi/ (kite), /ɑu/ (couch), /ɔi/ (boil), /ɝ/ (bird), /ə/ (about), /θ/ (thin), /ð/ (then), /š/ (shoe), /ž/ (beige), /ŋ/ (sing), /ʍ/ (which), /č/ (chin), /ǰ/ (jump).

stressed vowels, consonants in prestressed position, unstressed vowels, and consonants in poststressed (or unstressed) positions.

Stressed vowels are, in some sense, the sounds that are of most importance in an utterance, and it is almost essential that any speech-recognition scheme be able to locate and at least partially identify these sounds. Next in order of importance are the consonants and consonant clusters that precede stressed vowels. These might be regarded as the prototype consonants; the acoustic characteristics for consonants in this position should appear with clarity and should not be subject to the distortions and omissions that are typical of consonants in other phonetic environments. The vowels and consonants in unstressed positions are often greatly influenced by factors such as rate of talking and state of the talker. These sounds probably provide cues for recognition that are less reliable than those associated with vowels and consonants in stressed positions.

The utterances listed in Table II and generated by three talkers are not described exhaustively in this report. The purpose of the report is rather to examine some highlights of the data. An attempt is made to discuss all of the speech sounds, but not necessarily to consider the detailed characteristics of these sounds in all possible phonetic contexts that might be of interest. Most of the data presented here are for one speaker (KS), but examples from the other speakers are often shown either to corroborate the results for speaker KS or to indicate the kind of variability that might be expected from one speaker to another.

TABLE II. List of utterances used in phonetic study. The numbers are simply codes for identifying the utterances, particularly for purposes of computer analysis. Nonsense utterances are described in terms of phonetic symbols. Words are identified by their orthography.

No.	Utterance	No.	Utterance	No.	Utterance	No.	Utterance
101	ə'pip	125	ə'tɑt	149	ə'kɑk	173	ə'sos
102	ə'pɪp	126	ə'tʌt	150	ə'kʌk	174	ə'sɔs
103	ə'pep	127	ə'tɝ̈t	151	ə'kɝ̈k	175	ə'sæs
104	ə'pɛp	128	ə'did	152	ə'gig	176	ə'sɑs
105	ə'pup	129	ə'dɪd	153	ə'gag	177	ə'sʌs
106	ə'pʊp	130	ə'ded	154	ə'gug	178	ə'sɝ̈s
107	ə'pop	131	ə'dɛd	155	ə'fif	179	ə'ziz
108	ə'pɔp	132	ə'dud	156	ə'faf	180	ə'zɪz
109	ə'pæp	133	ə'dʊd	157	ə'fuf	181	ə'zez
110	ə'pɑp	134	ə'dod	158	ə'viv	182	ə'zɛz
111	ə'pʌp	135	ə'dɔd	159	ə'vav	183	ə'zuz
112	ə'pɝ̈p	136	ə'dæd	160	ə'vuv	184	ə'zʊz
113	ə'bib	137	ə'dɑd	161	ə'θiθ	185	ə'zoz
114	ə'bɑb	138	ə'dʌd	162	ə'θɑθ	186	ə'zɔz
115	ə'bub	139	ə'dɝ̈d	163	ə'θuθ	187	ə'zæz
116	ə'tit	140	ə'kik	164	ə'ðið	188	ə'zɑz
117	ə'tɪt	141	ə'kɪk	165	ə'ðɑð	189	ə'zʌz
118	ə'tet	142	ə'kek	166	ə'ðuð	190	ə'zɝ̈z
119	ə'tɛt	143	ə'kɛk	167	ə'sis	191	ə'v̈siv̈s
120	ə'tut	144	ə'kuk	168	ə'sɪs	192	ə'v̈sɑv̈s
121	ə'tʊt	145	ə'kʊk	169	ə'ses	193	ə'v̈sʊv̈s
122	ə'tot	146	ə'kok	170	ə'sɛs	194	ə'v̈siv̈z
123	ə'tɔt	147	ə'kɔk	171	ə'sus	195	ə'v̈zɑv̈z
124	ə'tæt	148	ə'kæk	172	ə'sʊs	196	ə'v̈zʊv̈z

TABLE II (*continued*)

No.	Utterance	No.	Utterance	No.	Utterance	No.	Utterance
197	ə'hi	224	ə'ji	251	ə'bra	278	sober
198	ə'ha	225	ə'ja	252	ə'bla	279	robin
199	ə'hu	226	ə'ju	253	ə'dra	280	rabid
200	ə'mim	227	ə'ɱi	254	ə'gla	281	baby
201	ə'mam	228	ə'ɱa	255	ə'skra	282	bottle
202	ə'mum	229	ə'ɱu	256	ə'spla	283	water
203	ə'nin	230	ə'ʧiʧ	257	bib	284	button
204	ə'nɪn	231	ə'ʧaʧ	258	bɪb	285	seated
205	ə'nen	232	ə'ʧuʧ	259	beb	286	detail
206	ə'nɛn	233	ə'ʤiʤ	260	bɛb	287	modal
207	ə'nun	234	ə'ʤaʤ	261	bæb	288	raider
208	ə'nʊn	235	ə'ʤuʤ	262	bab	289	hidden
209	ə'non	236	ə'spa	263	bɔb	290	beaded
210	ə'nɔn	237	ə'sta	264	bob	291	body
211	ə'næn	238	ə'ska	265	bʊb	292	local
212	ə'nan	239	ə'sma	266	bub	293	poker
213	ə'nʌn	240	ə'sna	267	bʌb	294	reckon
214	ə'nɝn	241	ə'kwa	268	bɝb	295	raucous
215	ə'lil	242	ə'twa	269	baɪb	296	cocoa
216	ə'lal	243	ə'pra	270	baʊb	297	legal
217	ə'lul	244	ə'tra	271	bɔɪb	298	sugar
218	ə'rir	245	ə'kra	272	apple	299	wagon
219	ə'rar	246	ə'pla	273	paper	300	ragged
220	ə'rur	247	ə'kla	274	open	301	pogo
221	ə'wiŋ	248	ə'sla	275	rapid	302	suffer
222	ə'waŋ	249	ə'swa	276	happy	303	muffin
223	ə'wuŋ	250	ə'fla	277	table	304	hovel

TABLE I (*continued*)

No.	Utterance	No.	Utterance	No.	Utterance	No.	Utterance
305	cover	318	mohair	331	flyer	344	bald
306	author	319	lemon	332	nowhere	345	heart
307	pathos	320	famous	333	kitchen	346	hard
308	father	321	dinner	334	ketchup	347	saunter
309	fathom	322	peanut	335	region	348	launder
310	lesser	323	single	336	edges	349	seltzer
311	essay	324	singer	337	gaunt	350	sudser
312	dozen	325	killer	338	bond	351	Gloucester
313	busy	326	pallid	339	lots	352	filter
314	bushel	327	horrid	340	sods	353	builder
315	nation	328	very	341	frost	354	martyr
316	measure	329	tower	342	fizzed	355	harder
317	vision	330	seaweed	343	fault		

4. STRESSED VOWELS

There are various schemes for classifying the vowels of American English. In our discussion here, we postulate that there are 15 vowels that can occur in stressed position. Three of these (/ɑi/ /ɑu/ and /ɔi/) are diphthongs, one (/ɝ/) is a retroflex vowel, and the remaining 11 can be categorized in terms of articulatory features in the manner shown in Table III.

TABLE III. Features of the vowels of American English. The symbol + indicates the presence of a feature, and the symbol - indicates the absence of a feature.

	i	ɪ	e	ɛ	æ	ɑ	ʌ	ɔ	o	ʊ	u
back	-	-	-	-	-	+	+	+	+	+	+
high	+	+	-	-	-	-	-	-	-	+	+
low	-	-	-	-	+	+	+	+	-	-	-
rounded	-	-	-	-	-	-	-	+	+	+	+
tense	+	-	+	-	-	+	-	+	+	-	+

The diphthongs /ɑi/, /ɑu/, and /ɔi/ can be classed as tense vowels. On the basis of acoustical data, we shall observe that certain vowels followed by /l/ and /r/ also have some of the characteristics of diphthongs.

In English there is a tendency for some of the tense vowels to be diphthongized; that is, the vowel quality changes with time through

the vowel. This diphthongization is particularly evident in the vowels /e/ and /o/, and is less apparent (but still observable) for the vowels /i/ and /u/. Thus, of the nine tense vowels, four terminate in an /i/-like position (/ɑi, ɔi, e, i/), three terminate in an /u/-like position (/ɑu, o, u/), and the two low back vowels (/ɑ/ and /ɔ/) are not diphthongized.

The lax vowels /ɪ, ɛ, æ , ʌ, ʊ/ are generally shorter than the tense vowels (although /æ/ may be an exception). Throughout their duration these vowels tend to drift toward an open or schwa vowel configuration (designated by /ə/), with the possible exception of /ʌ/, which is already close to that configuration. It is noteworthy that a (stressed) lax vowel is *always* followed by a consonant in English, whereas a tense vowel may appear in final position without a following consonant.

Some of the effects just noted can be observed in Fig. 4, which shows spectrograms of five of the vowels generated by one speaker in the environment /bVb/. For example, the diphthongization of /u/ is manifested in the falling second formant, while in /e/ the second and third formants are rising throughout the vowel.

Durations of the vowels in the environment /bVb/ are given in Table IV. These durations, which were measured from spectrograms, represent the time from release of the initial /b/ to the onset of the stop gap in the final /b/. It is evident that the tense vowels are the longest, with the exception of the lax vowel /æ/ and the retroflex vowel /ɝ/, which have durations comparable to those of tense vowels. More complete data on vowel durations for other consonantal environments in a nonsense-syllable frame have been reported by House (1961). As is well known, vowel durations

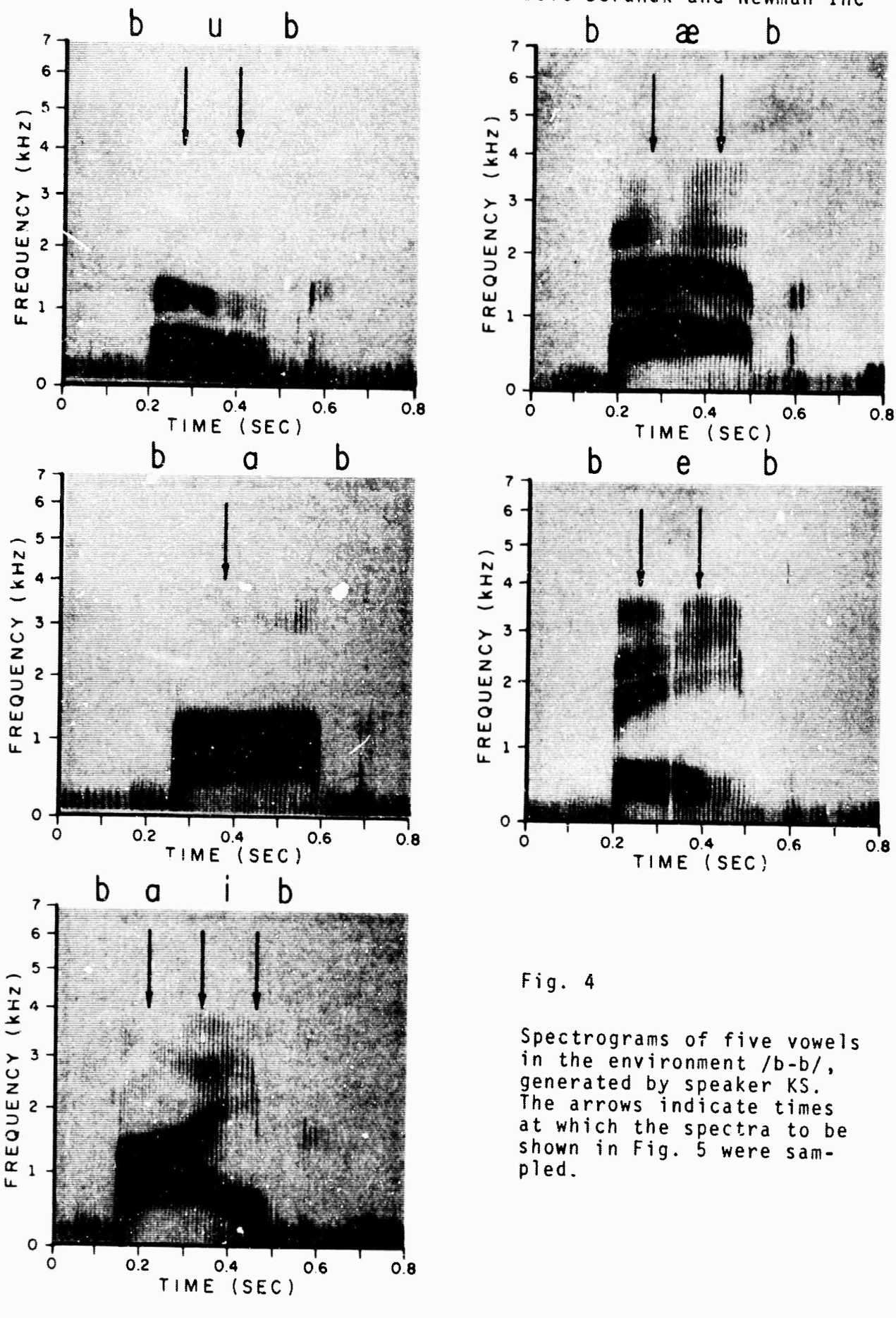


Fig. 4

Spectrograms of five vowels in the environment /b-b/, generated by speaker KS. The arrows indicate times at which the spectra to be shown in Fig. 5 were sampled.

depend to some extent on the following consonant, being longer for final voiced consonants and shorter for final voiceless consonants. The durations given in Table IV cannot be expected to remain invariant in utterances that are several syllables long. In natural speech, the vowel durations will usually be shorter, particularly when the stressed vowel is followed by a syllable containing an unstressed vowel, but the durations for different vowels will tend to maintain the same relative values as those shown in Table IV.

TABLE IV. Durations of stressed vowels in the environment /bVb/. Averages for three talkers generating one utterance for each vowel.

Vowel	Duration (msec)	Vowel	Duration (msec)	Vowel	Duration (msec)
i	300	ɑu	330	ɪ	170
e	300	ɑi	320	ɛ	200
ɑ	330	ɔi	280	æ	330
ɔ	310	ɝ	270	ʌ	180
o	290			ʊ	170
u	260				

The durations of the stressed vowels in some of the bisyllabic words (with stress on the initial syllable) were also measured. As noted earlier, the words included some consonant clusters in syllable-final position. For the clusters containing /l/ and /r/ (as in the words *filter* and *harder*), it is not possible to

establish a boundary between the vowel and the consonant for purposes of duration measurements. For such utterances, therefore, durations of the entire vowel-consonant combination (e.g., /ɪl/ and /ɑr/) were measured. The results of these measurements are shown in Table V. For purposes of these measurements, the onset of the vowel is considered to be at the instant of consonant release. Individual data for each of the three speakers are given in order to provide an indication of the variability to be expected from speaker to speaker. The words are arranged in order of increasing mean vowel duration.

Comparison of the vowel durations of single stressed vowels in bisyllabic words (in Table V) with the corresponding vowels in the monosyllabic utterances /bVb/ from Table IV indicates that the duration of a stressed vowel is often considerably shortened in a multisyllabic utterance. For some vowels, the duration in the bisyllabic context is as little as one-third of the duration in a monosyllable. When the vowel /ɑ/ is followed by /r/, or when /ɪ/ or /ɛ/ are followed by /l/, the total duration of the vowel-sonorant combination is comparable to that of a tense vowel when that vowel is followed by an obstruent (i.e., a consonant produced with complete closure or with noise at the constriction). Thus there is a tendency for these vowel-sonorant clusters to behave like tense vowels or diphthongs as far as their durations are concerned.

In general, then, the durations of stressed vowels, diphthongs, or vowel-sonorant combinations may be as short as 80 msec and as long as 380 msec. The shorter vowels are the single lax vowels in bisyllabic words with stress on the first syllable, while the longer ones are tense vowels or diphthongs in monosyllabic isolated words.

TABLE V. Measured vowel durations (in milliseconds) from vowels and vowel-consonant combinations occurring in bisyllabic words spoken in isolation. The underscored items indicate the segments whose durations were measured. The words are arranged in order of increasing duration value.

	Speaker			Mean
	KS	CW	GC	
sugar	82	82	83	82
hidden	83	75	98	85
seated	98	105	128	110
suffer	112	112	105	110
button	120	120	90	110
busy	112	135	143	130
ketchup	135	142	113	130
Gloucester	142	150	158	150
seltzer	150	165	143	153
filter	120	173	188	160
saunter	158	158	165	160
rapid	150	158	188	165
paper	158	172	165	165
bottle	173	188	158	173
martyr	165	173	195	178
harder	165	173	203	180
modal	165	195	188	183
raucous	202	188	165	185
water	188	195	180	188
launder	202	158	225	195
robin	195	202	240	212
builder	217	240	240	232
father	240	225	263	243

A comprehensive model that accounts for the variation in duration for stressed vowels in various environments (including longer words and phrases, which are not examined in this study) has yet to be developed. The duration of a vowel in a multisyllabic utterance is evidently influenced by the "rhythm" of the utterance, including the timing between syllabic nuclei, and the factors that determine this timing of these gross aspects of an utterance are not known at present. Informal observations indicate, however, that in a multisyllabic utterance the time intervals between vowels with stress are much less variable than the durations of individual stressed vowels.

Spectra of the 15 vowels uttered in the context /bVb/ by one of the speakers are shown in Fig. 5. These spectra are actually smoothed outputs of the 19-channel filter bank referred to earlier. The solid curves for each of these vowels represent spectra taken at about 70–100 msec after the release of the initial /b/. In cases where the vowels appear to be diphthongized, or to drift towards a schwa configuration, one or more additional spectra are shown. These are samples at later instants of time in the vowels. For each vowel, the spectrum samples are identified by number. These numbers simply designate which 10-msec interval was examined, and the numbers begin at an arbitrary point just prior to the onset of the utterance, as shown in the sample of the printout displayed in Fig. 3.

For four of the vowel spectra (/i, æ, ɑ, u/), arrows are drawn to indicate the frequencies of the lowest two or three formants as measured from the spectrograms.* It is evident that the spectra

*The formant frequencies for these and other vowels are in the ranges reported by Peterson and Barney (1952), who did an exhaustive study of the formant frequencies of a number of vowels in the context hVd, spoken by many different people.

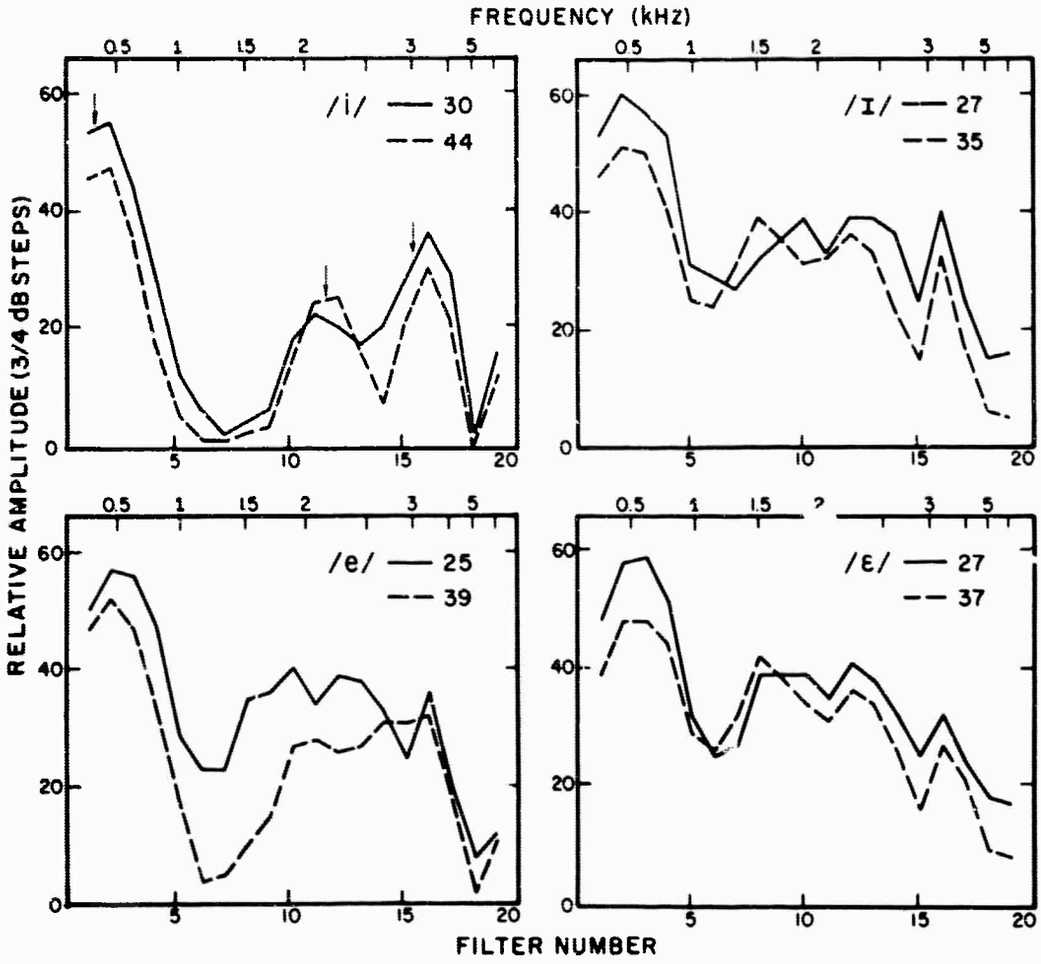


Fig. 5 Spectra of 15 stressed vowels in the environment /b-b/ obtained from 19-channel filter bank Curves are labeled with sample number (10-msec sampling interval) beginning at an arbitrary time prior to onset of utterance. Solid lines represent spectra sampled 70–100 msec after onset of initial consonant. Dashed lines represent spectra sampled later in the vowel for cases where there is an appreciable shift in spectrum. Spectra are sampled at three points throughout the diphthongs. For the vowels /i/ /æ/ /ɑ/ /u/, the small arrows indicate the frequencies of formants as measured from spectrograms. Data are for speaker KS.

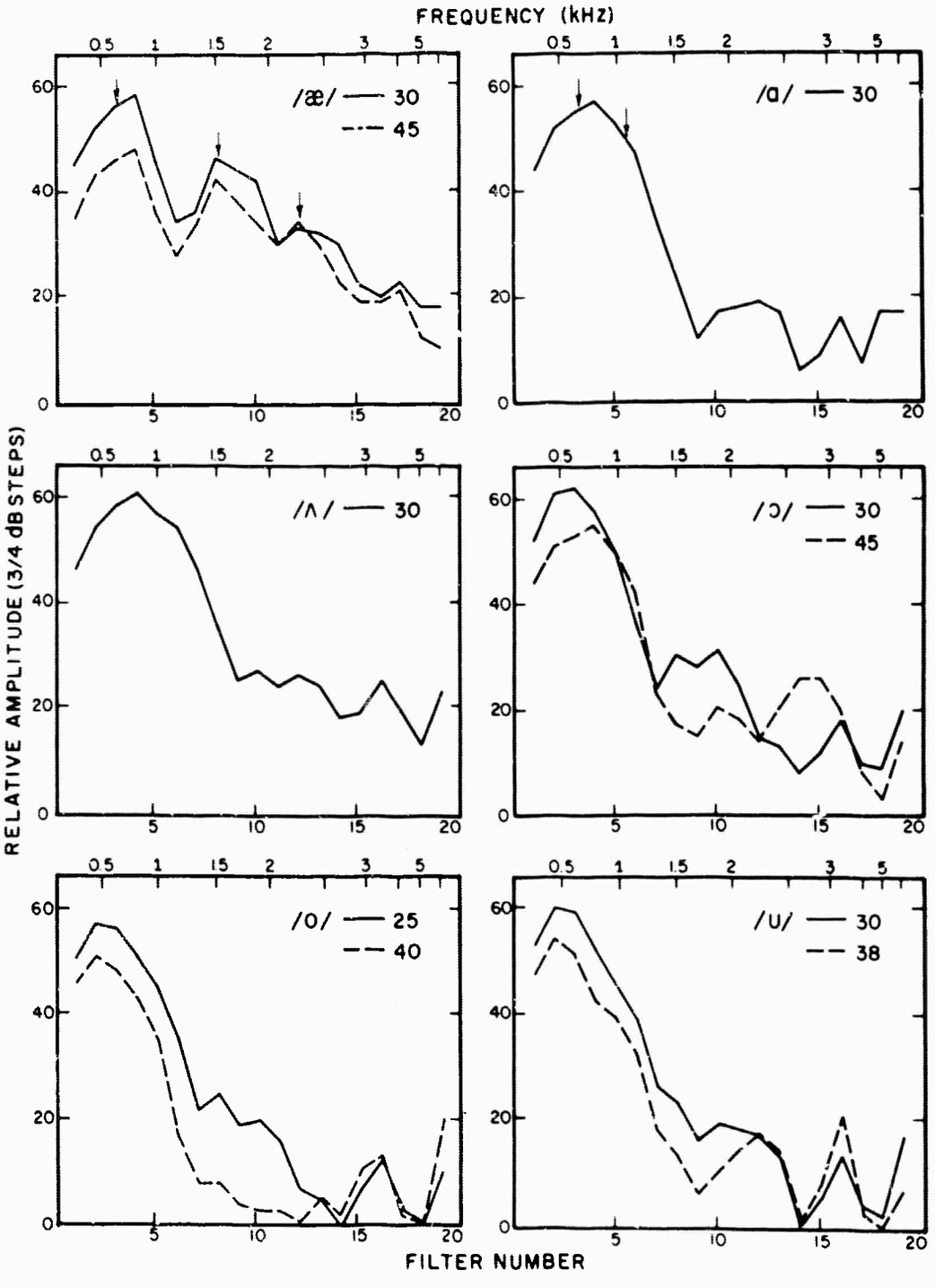

FREQUENCY (kHz)
0.5 1 1.5 2 3 5
/æ/ — 30
-·- 45
/ɑ/ — 30
/ʌ/ — 30
/ɔ/ — 30
-- 45
/o/ — 25
-- 40
/u/ — 30
-- 38
RELATIVE AMPLITUDE (3/4 dB STEPS)
0 20 40 60
5 10 15 20
FILTER NUMBER

Fig. 5 (Cont.)

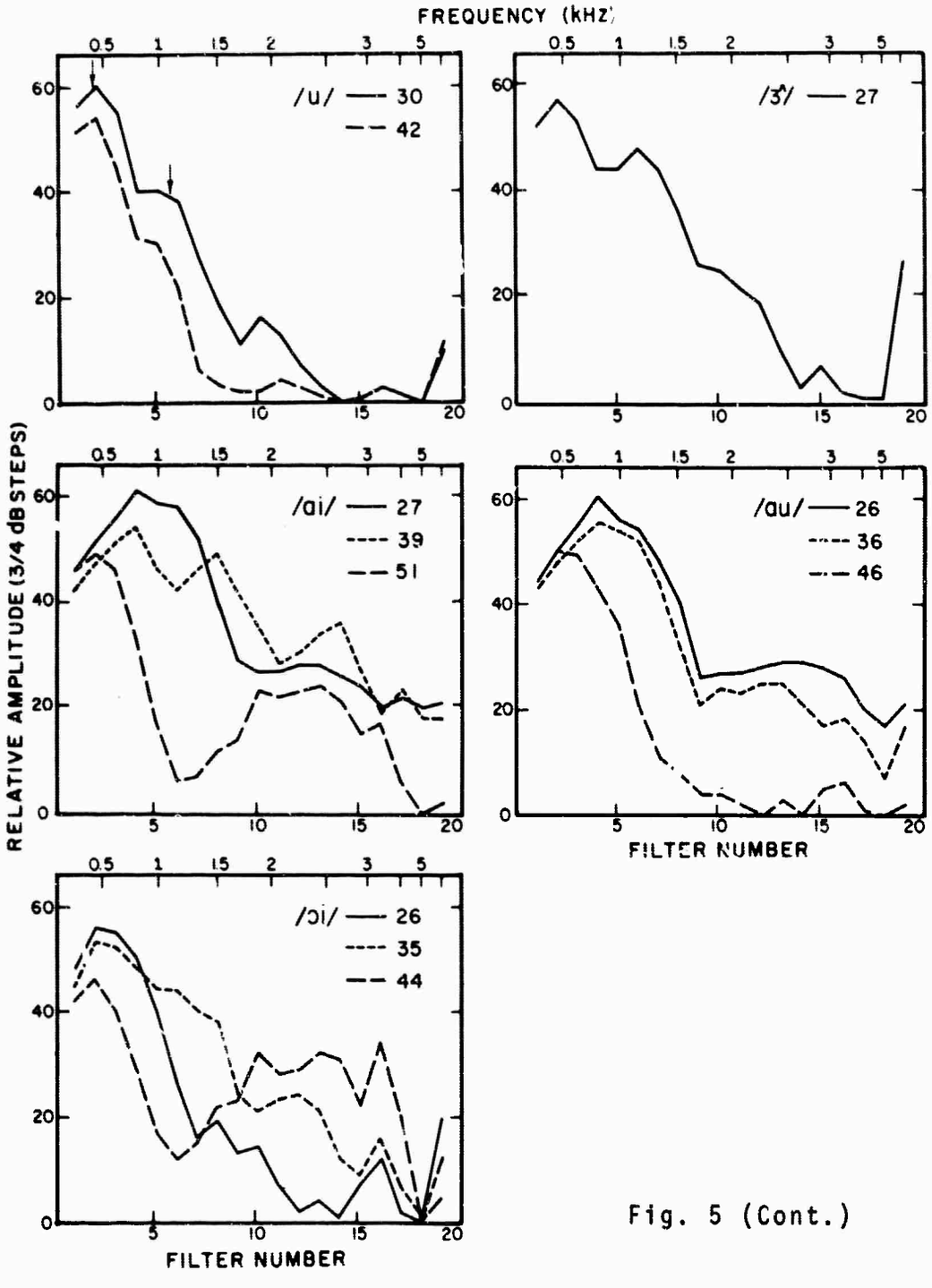

FREQUENCY (kHz)
0.5 1 1.5 2 3 5
/u/ — 30
-- 42
/ɜ/ — 27
/ai/ — 27
---- 39
-- 51
/au/ — 26
---- 36
-·- 46
/ɔi/ — 26
---- 35
-- 44
RELATIVE AMPLITUDE (3/4 dB STEPS)
FILTER NUMBER
0 20 40 60
5 10 15 20

Fig. 5 (Cont.)

obtained from the relatively broad filters in the 19-channel filter bank have peaks in the vicinity of formant frequencies. However, when two formants are sufficiently close together, the spectrum representation from the filter bank may show only a single broad peak. We regard this poor resolution of spectral peaks not to be a drawback in the 19-channel spectrum analyzer; there is some evidence (Fant, 1959; Fujimura, 1967) that two closely spaced formants tend to be interpreted perceptually in the same way as a single energy concentration, whether the two formants are F1 and F2 (Formant 1 and Formant 2), as for back vowels, or F2 and F3, as for front vowels. Furthermore, theoretical considerations show that the relative amplitudes of different regions of vowel spectra are dependent on the formant frequencies,* and hence provide information regarding the locations of formants (Fant, 1956; Stevens and House, 1961).

Several gross properties of the data are apparent from Fig. 5. All the vowels are characterized by a major energy concentration at low frequencies, the peak being in one of the three filters 2, 3, and 4 (spanning the frequency range 260 to 980 Hz). This region corresponds, of course, to the first-formant frequency, although in the case of back vowels /ɑ, ɔ, ʌ, o, ʊ, u/ the low-frequency peak is a consequence of both the first and the second formants, which are close together for these vowels.

The front vowels (i, ɪ, e, ɛ, æ) always have one or more additional energy peaks at high frequencies (filter number 8 – 1520 Hz – or

*These amplitude relations indicate that an increase in the frequency of a given formant causes an increase in the amplitude of the spectrum peaks corresponding to formants located at frequencies above that formant. Also, when two formants move close together in frequency, the amplitude of the spectrum in the vicinity of the frequencies of these formants increases.

higher) and a significant energy minimum between the low- and high-frequency peaks. This energy minimum for this speaker is always in the range of filters 6 and 7 (1160 to 1340 Hz), and the minimum value is always at least 10 units (7.5 dB) below the higher frequency peak. The amplitude of the high-frequency filter with maximum energy is no more than 20 units (15 dB) below the peak amplitude at low frequencies for this speaker.

The back vowels have no such deep valley in the spectrum. If an energy minimum exists in this frequency range (filters 6 and 7), it is a rather shallow minimum, and the peak amplitude of the higher-frequency peak is well over 20 units below that of the low-frequency peak.

The various front vowels are distinguished from one another by the position of the low-frequency energy peak and by the distance (in frequency) between the central energy minimum and the adjacent high- and low-frequency peaks. For /i/, the low peak is at filter 2 (440 Hz) and the higher-frequency concentration is at filter 10 (1880 Hz) and above. At the other extreme, the low front vowel /æ/ has the low-frequency peak at filter 4 (800 Hz) and the higher peak at filter 8 (1520 Hz).

Distinctions among the various back vowels are made primarily on the basis of the frequency width and position of the major low-frequency energy concentration, which is a consequence of the first and second formants. This energy concentration is lowest in frequency for /u/ and highest in frequency for /ɑ/ and /ʌ/, with /ʊ ɔ/ lying between. The high frequencies (above filter 9 at 1700 Hz) are sufficiently weak that they do not play an important role in distinguishing between the back vowels.

The diphthongized vowel /e/ moves from a spectrum shape intermediate between /i/ and /æ/ toward an /i/-like configuration. The vowel /o/, on the other hand, has an initial spectrum shape intermediate between /u/ and /ɑ/ and then glides toward a /u/-like configuration. These diphthongs in this phonetic context of an isolated CVC utterance are characterized by a decrease in amplitude of the low-frequency peak as the glide proceeds toward the extreme high position characteristic of the /i/ or /u/. Likewise, the vowels /i/ and /u/ show some diphthongization toward more-extreme configurations; for both vowels there is a slight drop in frequency of the low-frequency peak as well as a decrease in its amplitude.

The diphthongs /ɑi, ɑu, ɔi/ show the expected motions between two vowel configurations. For /ɑi/, the spectrum near the beginning is like that of the vowel /ɑ/; it moves, at first slowly and then more rapidly, toward an /i/ or /ɪ/ spectrum. The most obvious effect for this diphthong is the introduction of the midfrequency minimum as the vowel glides from a back configuration to a front configuration. The combination /ɔi/ has similar characteristics. The diphthong /ɑu/ is, of course, a back vowel throughout its length, and the movement is primarily a shift of the low-frequency peak in the downward directon, with a resulting decrease in amplitude in the high-frequency range.

The acoustic data for the tense vowels and diphthongs provide evidence, therefore, that one set of vowels (/i, e, ɑi, ɔi/) is diphthongized with a final glide toward an /i/-like spectrum, whereas another set (/u, o, ɑu/) has a final glide toward a /u/-like spectrum, as discussed earlier. The low vowels /ɑ/ and /ɔ/ are the only tense vowels in American English that do not have one of these two glides.

The lax vowels /ɪ, ɛ, æ, ʊ/ appear also to exhibit a change in spectrum as a function of time. For each of these vowels, the spectrum towards the end of the vowel tends to have a second-formant peak in the vicinity of filter 8 (1520 Hz). (For /ʊ/, and to some extent for /ʌ/, this tendency is more apparent in utterances with the final consonant /d/; the final consonant always has some influence on the vowel spectrum near the end of the vowel.) Such a second-formant frequency is characteristic of the schwa vowel /ə/.

Thus any stressed vowel in which there is a drift toward the schwa position must be a lax vowel. No tense vowel has this property. It may be significant also that there is a reduction in amplitude of the low-frequency peak (by about 10 units) during the drift toward the schwa position, but there is no appreciable drop in amplitude of the high frequency peak. In the /ɛ/, for example, the low peak decreases in amplitude by 11 units, and shifts downward slightly in frequency, whereas the higher peak changes very little in amplitude.

By the criteria discussed above, /ɝ/ would be classified as a back vowel since it does not have a pronounced midfrequency minimum. The secondary peak at filter 6, which is of appreciable amplitude relative to the low-frequency maximum for this vowel, serves to distinguish /ɝ/ from the other back vowels. The relatively high amplitude of this peak is presumably due to the proximity of the second and third formants in the frequency range 1000–1600 Hz.

The vowel spectra for the other two speakers exhibit characteristics similar to those shown in Fig. 5. Spectra of three of the

the vowels for the three speakers are displayed in Fig. 6. In all cases, the vowels were in the environment /bVb/; the spectra were sampled 70–100 msec following the release and again at a later time in cases where the spectrum changed appreciably. In the case of the vowel /i/, all spectra have a low- and a high-frequency energy concentration, with a broad valley between these peaks. There is a slight diphthongization toward a lower first-formant frequency and a higher second-formant frequency for all speakers, but there are appreciable differences in the shape of the high-frequency spectrum (above about 2000 Hz). For the vowel /ɪ/, the shift toward a schwa vowel (higher first formant, lower second formant) is evident for all speakers, but is more pronounced for some speakers than for others. The vowel /ɑ/ has the broad low-frequency peak for each of the three speakers.

The vowel spectra shown in Figs. 5 and 6 are influenced to some extent by the phonetic environment in which the stressed vowels occur. Effects of consonantal environment on vowel formant frequencies have been shown previously (Lehiste and Peterson, 1961; Stevens and House, 1963), and indicate that adjacent consonants tend to influence lax vowels more than tense vowels. An illustration of the effect of phonetic environment for a lax vowel is given in Fig. 7, which illustrates the range of spectra observed in the middle of the stressed vowel /ɪ/ in seven different nonsense syllables and bisyllabic words. The range tends to be greater at high frequencies than at low frequencies, presumably since small shifts in a formant frequency have a greater influence on spectrum amplitudes above that frequency than below it. Two examples of more deviant spectra of /ɪ/ are also shown in Fig. 7. The following consonants /ŋ/ and /l/ appear to have a strong effect on this vowel.

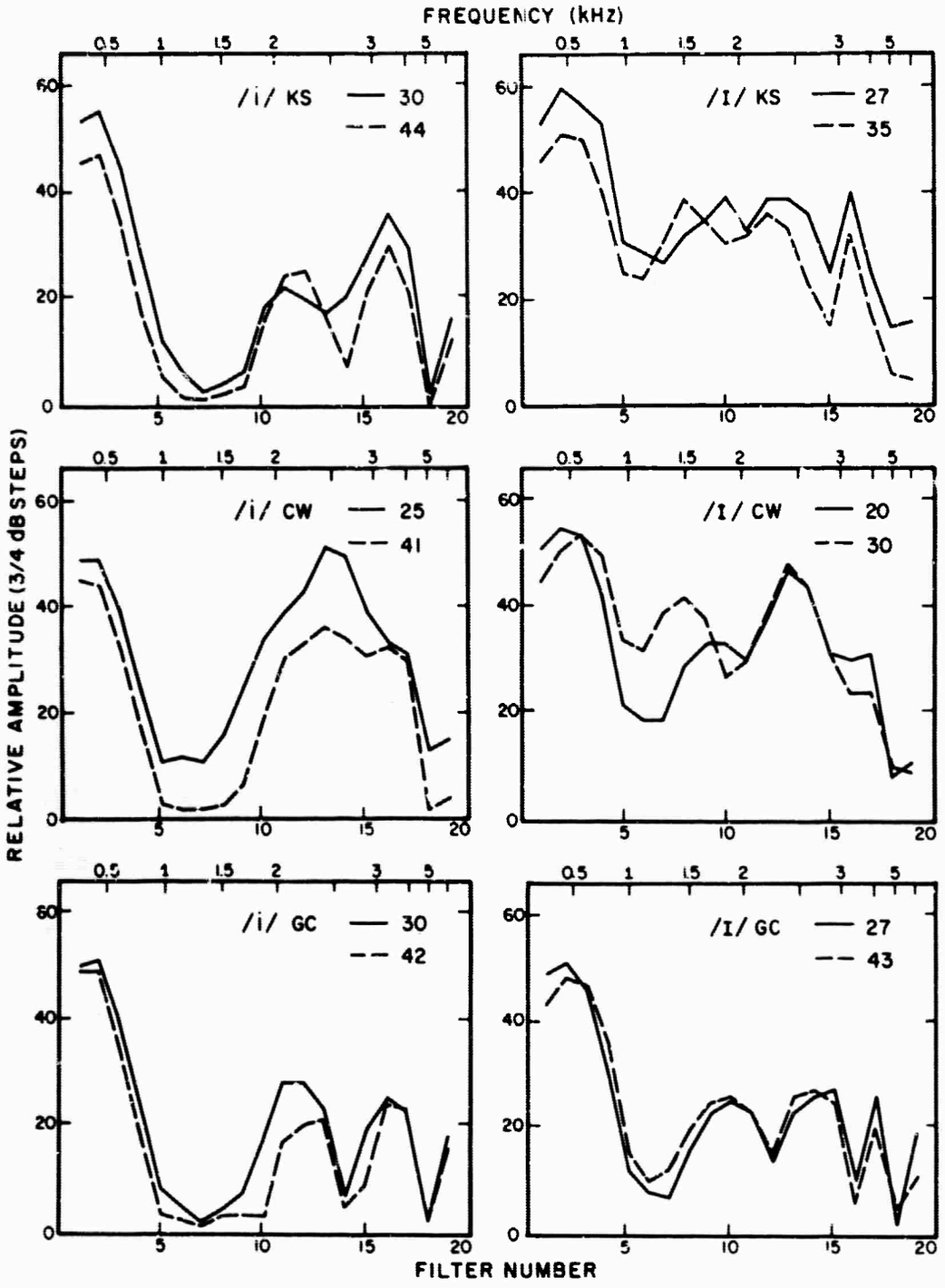


Fig. 6 Spectra of three stressed vowels (in the environment /b-b/) are compared for three speakers. Data obtained from 19-channel filter bank. See legend of Fig. 5.

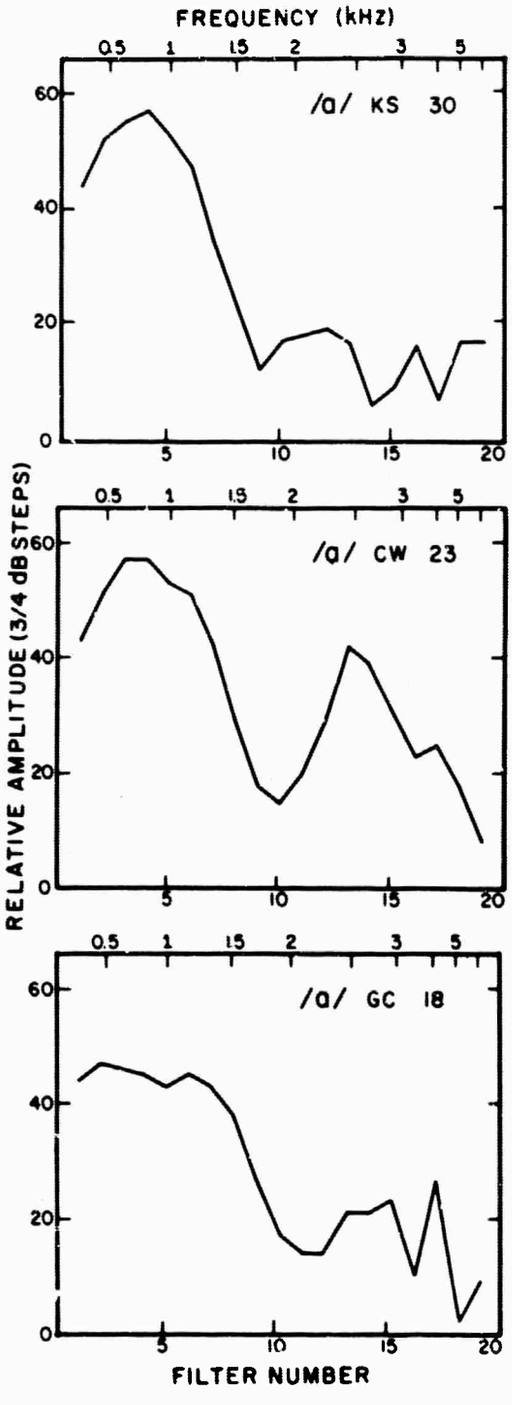

FREQUENCY (kHz)
0.5 1 1.5 2 3 5
/a/ KS 30
/a/ CW 23
/a/ GC 18
RELATIVE AMPLITUDE (3/4 dB STEPS)
0 20 40 60
5 10 15 20
FILTER NUMBER

Fig. 6 (Cont.)

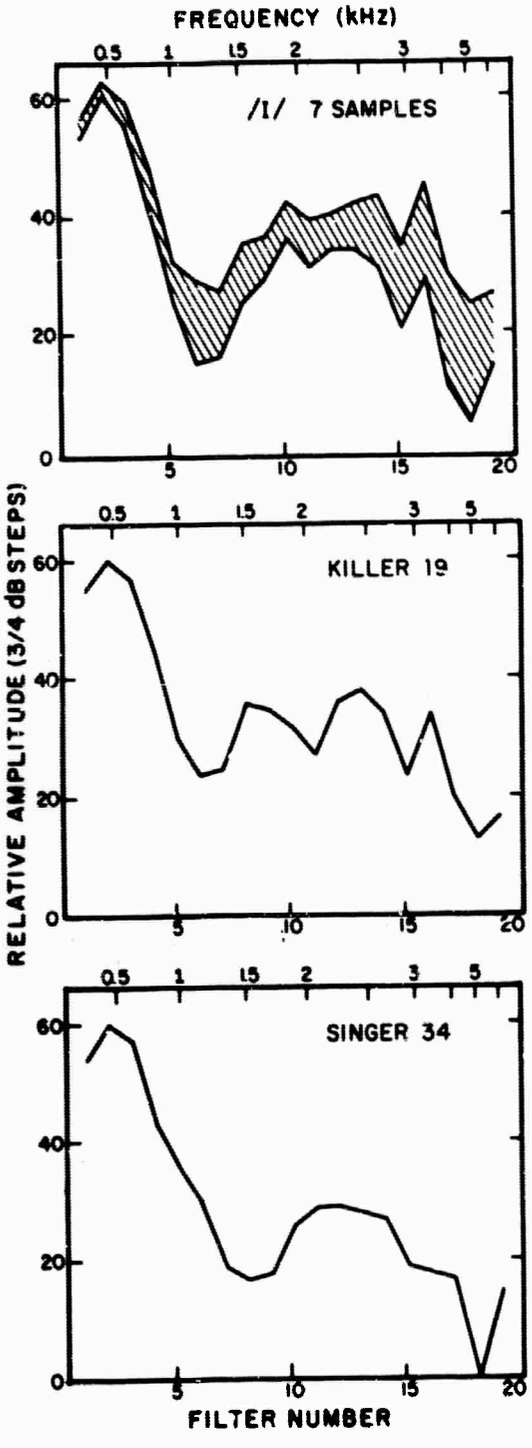


Fig. 7

The upper graph shows the range of spectra for the stressed vowel /ɪ/ for seven different consonantal environments in nonsense syllables and in bisyllabic words. The lower two graphs are examples of spectra of the same vowel when it is modified appreciably by the final consonant. Speaker KS.

The data of the type shown in Figs. 5 – 7 and in Tables IV and V suggest the possibility of developing algorithms that would be useful in separating one class of vowels from another. Although the purpose of this report is not to present such algorithms, it may be of interest to suggest some possibilities in order to indicate the kind of results that might be expected. Consider, for example, the separation of vowels into the classes *front* and *back*. (The diphthongs are omitted for the purposes of this analysis.) For the speakers examined in this study, front vowels are always characterized by a spectral minimum in the range of filters 6 and 7 (980–1520 Hz), and the low- and high-frequency spectral maxima are roughly at equal distances (in hertz) on either side of this minimum. For back vowels, the high-frequency spectral maximum, if it exists, is of much lower amplitude than the corresponding maximum for front vowels.

An algorithm that would roughly take these facts into account is the following:

(1) Look at the outputs of filters 5 through 8. If a minimum does not occur in filters 6 and 7 in this region, the vowel is a back vowel. (This procedure identifies all but a few of the back vowels examined in the /bVb/ utterances of this study.) If such a minimum does occur, record its value A_m in filter a.

(2) Find the spectral maximum below filter a. Say it is in filter b. Find filter c an equal distance above a; i.e., $c-a = a-b$. Determine the maximum A_n in one of the three filters $c \pm 1$. Compute A_n-A_m. If this difference is less than 8 units (about 6 dB), then the vowel is a back vowel. Otherwise it is a front vowel.

Examination of a number of the vowels (all the vowels in the /bVb/ context and a number of others) reveals that this algorithm divides the stressed vowels into two classes as desired, with no overlap.

Similar algorithms could be developed for other vowel features. The features *high* and *low*, for example, would be identified, at least in part, by the position of the low-frequency peak.

In summary, then, the attributes that distinguish one stressed vowel from another must include (1) duration, (2) spectrum characteristics, and (3) how the spectrum changes with time. The kind of display provided by the 19-channel filter bank seems to contain enough information to permit the stressed vowels to be distinguished from one another using these types of criteria.

5. CONSONANTS IN PRESTRESSED POSITION: SINGLE CONSONANTS

There are about 23 consonants that can appear in prestressed position in American English. One procedure for classifying these consonants in terms of binary features is given in Table VI. This is a slightly modified version of the system proposed by Chomsky and Halle (1968). Place of articulation for the consonants in this system is specified by the features *anterior* and *coronal*. The feature *anterior* indicates that the consonant is generated in front of the palato-alveolar region of the mouth, and *coronal* designates a consonant generated with the blade of the tongue. The term *aspiration* applies to a consonant for which noise energy is generated at the glottal opening following the consonantal release. A *sonorant* consonant is generated with no major obstruction to the air flow above the vocal cords. In discussing the acoustic characteristics of various consonants, the sounds will be grouped roughly into classes suggested by the feature description of Table VI. Other features can be used to characterize certain of these consonants, but those features will not be referred to in this report.

The acoustic information necessary for the identification of these consonants is of two kinds: (1) the characteristics during the constricted consonant interval preceding the release into the vowel; and (2) acoustic events at the release of the constricted interval and during the 50- to 100-msec interval in which there is a transition into the vowel. For consonants in intervocalic or in final position, the transition from the preceding vowel into the consonant also provides information about the consonant. We consider first some data obtained within the constricted consonant interval.

TABLE VI. Classifications of consonants in terms of distinctive features.

	p	t	k	b	d	g	f	θ	s	š	v	ð	z	ž	č	j	m	n	ŋ	l	r	w	j
Stop	+	+	+	+	+	+	-	-	-	-	-	-	-	-	+	+	+	+	+	+	-	-	-
Sonorant	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	+	+	+	+	+
Anterior	+	+	-	+	+	-	+	+	+	-	+	+	+	-	-	-	+	+	-	+	-	+	-
Coronal	-	+	-	-	+	-	-	+	+	+	-	+	+	+	+	+	-	+	-	+	+	-	-
Voiced	-	-	-	+	+	+	-	-	-	-	+	+	+	+	-	+	+	+	+	+	+	+	+
Nasal	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	+	-	-	-	-
Aspiration	+	+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

5.1 Closure interval: stop and nasal consonants

The stop and nasal consonants can be roughly placed in the same class, since they all have a reasonably steady-state interval followed by a discontinuous change at the instant of release into the following vowel. Spectrograms giving examples of each class of consonants for one speaker are shown in Fig. 8.

During the voiceless stop consonants, the constricted interval is silent, whereas in voiced stops there may be some vocal-cord vibration within this interval. The filter bank gives spectra of the form shown in Fig. 9 during this voicing interval. There is energy essentially only in the lowest two frequency bands, and the amplitude in the lowest filter is 20–30 units below the peak amplitude (maximum amplitude in any one of filters 2, 3, 4) during the following stressed vowel. The spectrum is more or less the same, independent of which stop consonant is involved, but the overall amplitude depends to some extent upon the speaker, the consonant, and the following vowel.

For nasal consonants, the spectrum within the closure interval is of higher intensity and is characterized by relatively greater energy at higher frequencies, as shown in Fig. 10. The spectral maximum is in filter 1 or in filter 2, and is usually about 10–15 units below the peak amplitude in the following vowel. There are no significant and consistent differences in the spectra for /m/ and for /n/, and the following vowel does not have an appreciable effect on the spectrum, at least when it is displayed in this relatively gross manner. All nasal consonants appear to have relatively weak spectral energy in the vicinity of filter 4 (800 Hz) relative to the energy at lower frequencies (Fujimura, 1962).

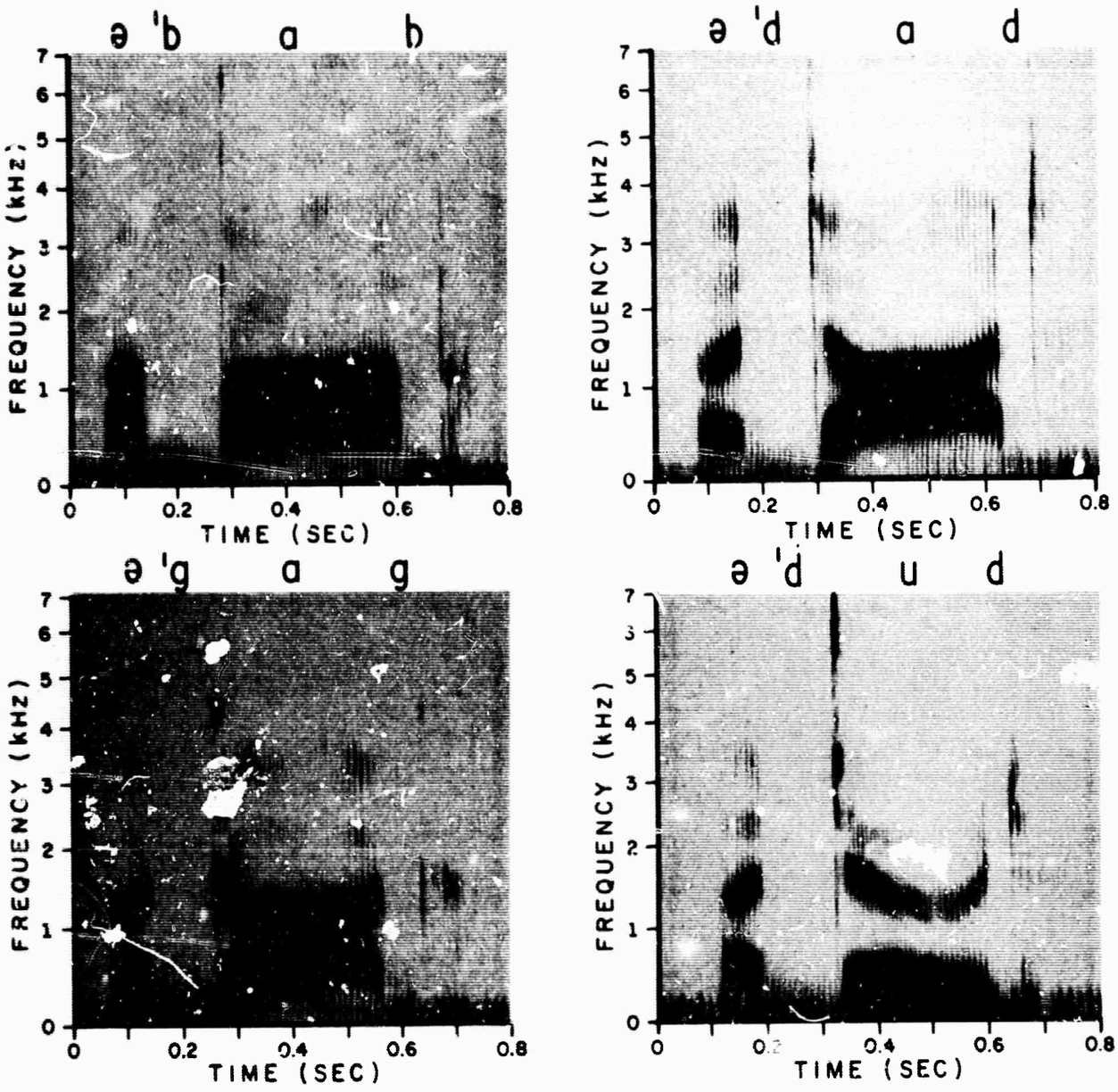
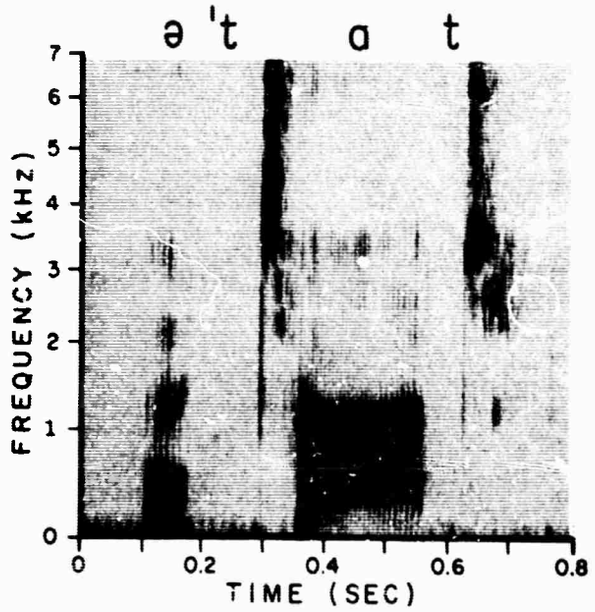
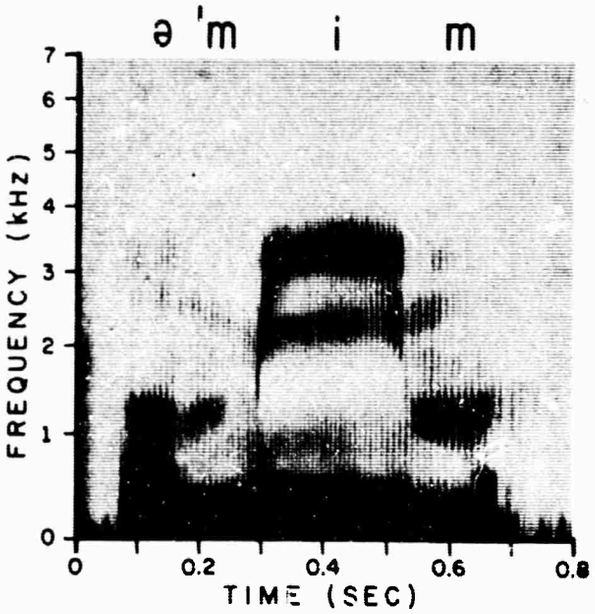


Fig. 8 Spectrograms illustrating properties of stop and nasal consonants. Speaker KS.


ə 't a t
FREQUENCY (kHz)
TIME (SEC)


ə 'm i m
FREQUENCY (kHz)
TIME (SEC)

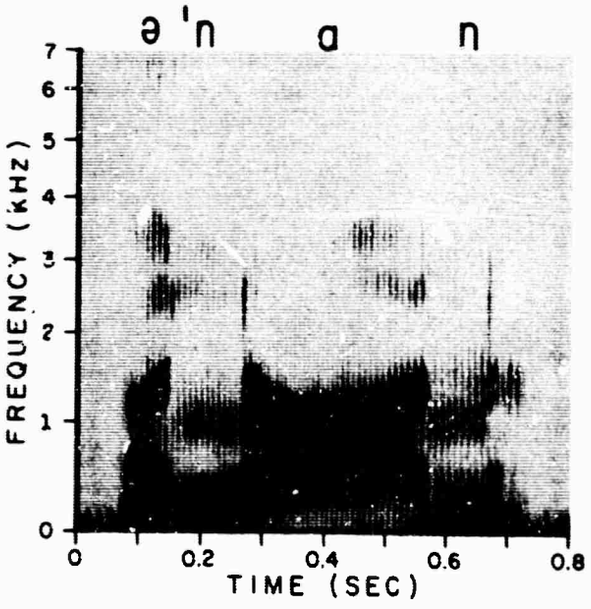

ə 'n a n
FREQUENCY (kHz)
TIME (SEC)

Fig. 8 (Cont.)

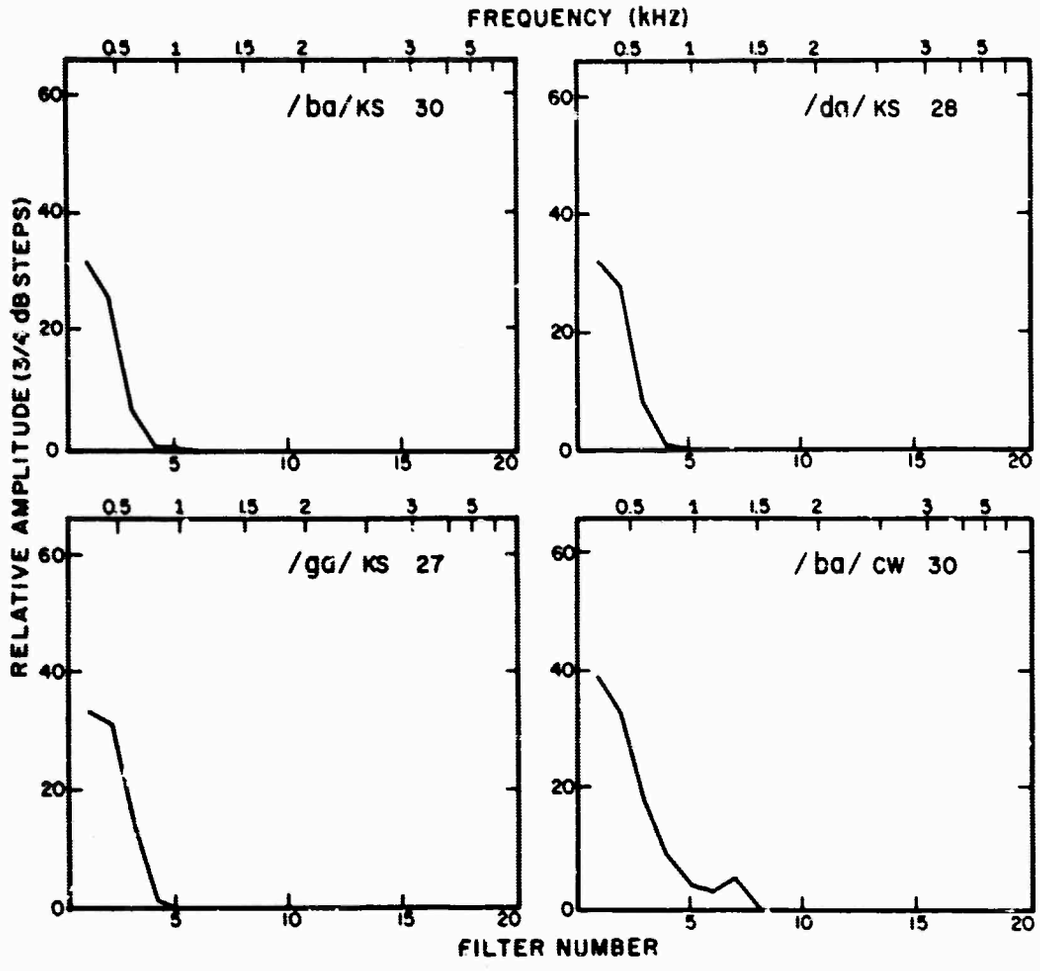


Fig. 9 Spectra during closure interval for voiced stop consonants obtained from 19-channel filter bank.

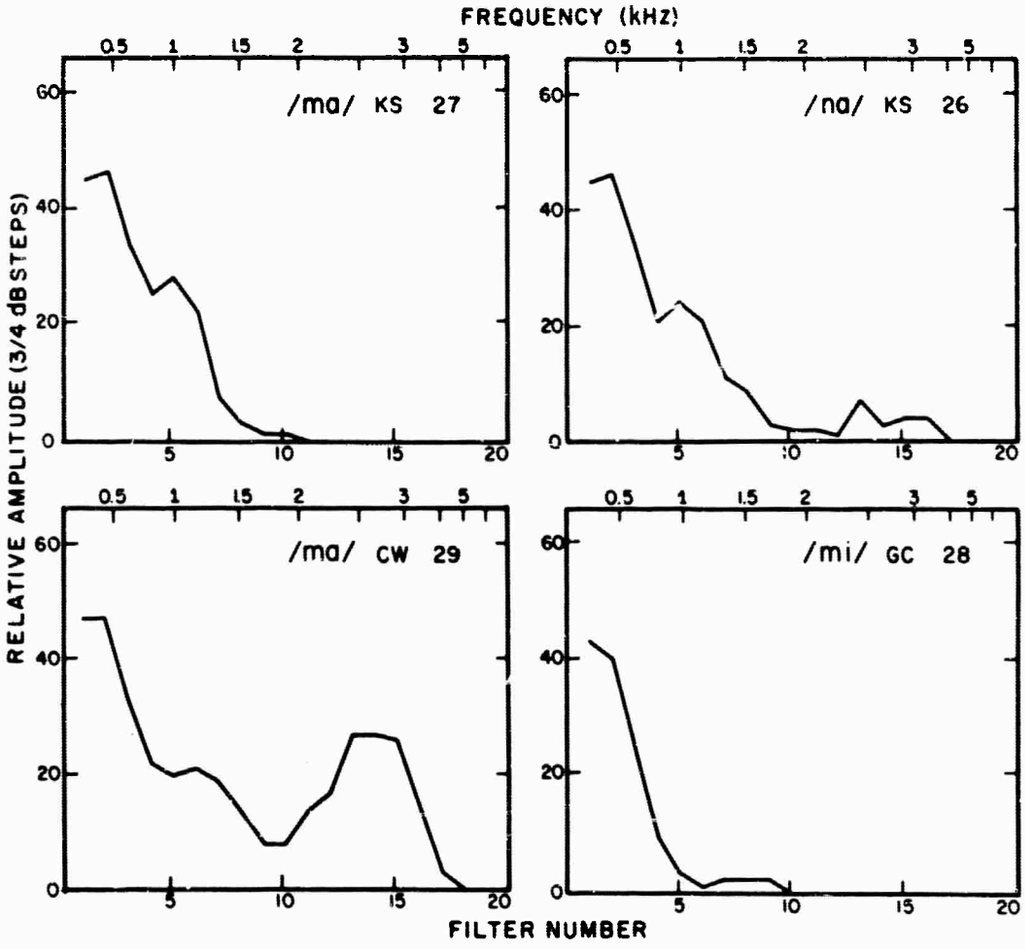


Fig. 10 Spectra during closure for nasal consonants in prestressed position.

This attribute distinguishes nasal consonants from /l/, as noted later. The detailed shape of the spectra for the nasals at high frequencies may differ considerably from one speaker to another, as comparison of /m/ for CW and GC in Fig. 10 demonstrates.

TABLE VII. Durations of closure intervals for stop and nasal consonants preceding stressed vowels. Averages over three vowel environments /i ɑ u/ and over three talkers.

Consonant	Average Duration (msec)
p	130
t	120
k	110
b	130
d	120
g	120
$\overset{v}{c}$	110
$\overset{v}{j}$	110
m	120
n	130
l	130

Durations of the closure intervals for stop and nasal consonants in the environment /ə'CV/ have been measured from spectrograms. These durations, averaged over three vowel environments, are listed in Table VII. There are no significant effects of vowel

environment, and there are only slight differences in duration between the three classes of consonants (voiceless stops, voiced stops, and nasals). Individual utterances may have stop gap durations that differ from the average values by as much as 25 percent. The average duration for the /l/ closure is also shown in Table VII, since in many respects /l/ can be classified with stops and nasals. These differences in duration between the various classes of consonants are much more marked for the consonants in poststressed position, as will be observed later. Likewise, the durations of stop gaps in prestressed position may be considerably shorter (by as much as 50 percent) than the values given in Table VII when the consonants are generated in the context of a longer speech sample.

5.2 Constricted interval: fricative consonants

The acoustic spectra within voiceless fricative consonants are always characterized by high-frequency energy, although in the case of the consonants /f/ and /θ/ this energy may be weak (Hughes and Halle, 1956; Heinz and Stevens, 1961). Spectrograms of the four voiceless fricative consonants preceding the vowel /ɑ/ are shown in Fig. 11. Typical spectra of these consonants are plotted in Fig. 12. Spectra for the other two speakers have similar gross characteristics. As is well known, the lowest major energy concentration in the spectrum for /š/ is in the frequency range 2000–3000 Hz (filters 13 to 14 in the example shown in Fig. 12), and there are further energy peaks at still higher frequencies. For /s/, on the other hand, the increase in spectral energy does not begin until filter 16–17 (3560–4400 Hz). The consonants /f/ and /θ/ have some high-frequency energy only in filter 19

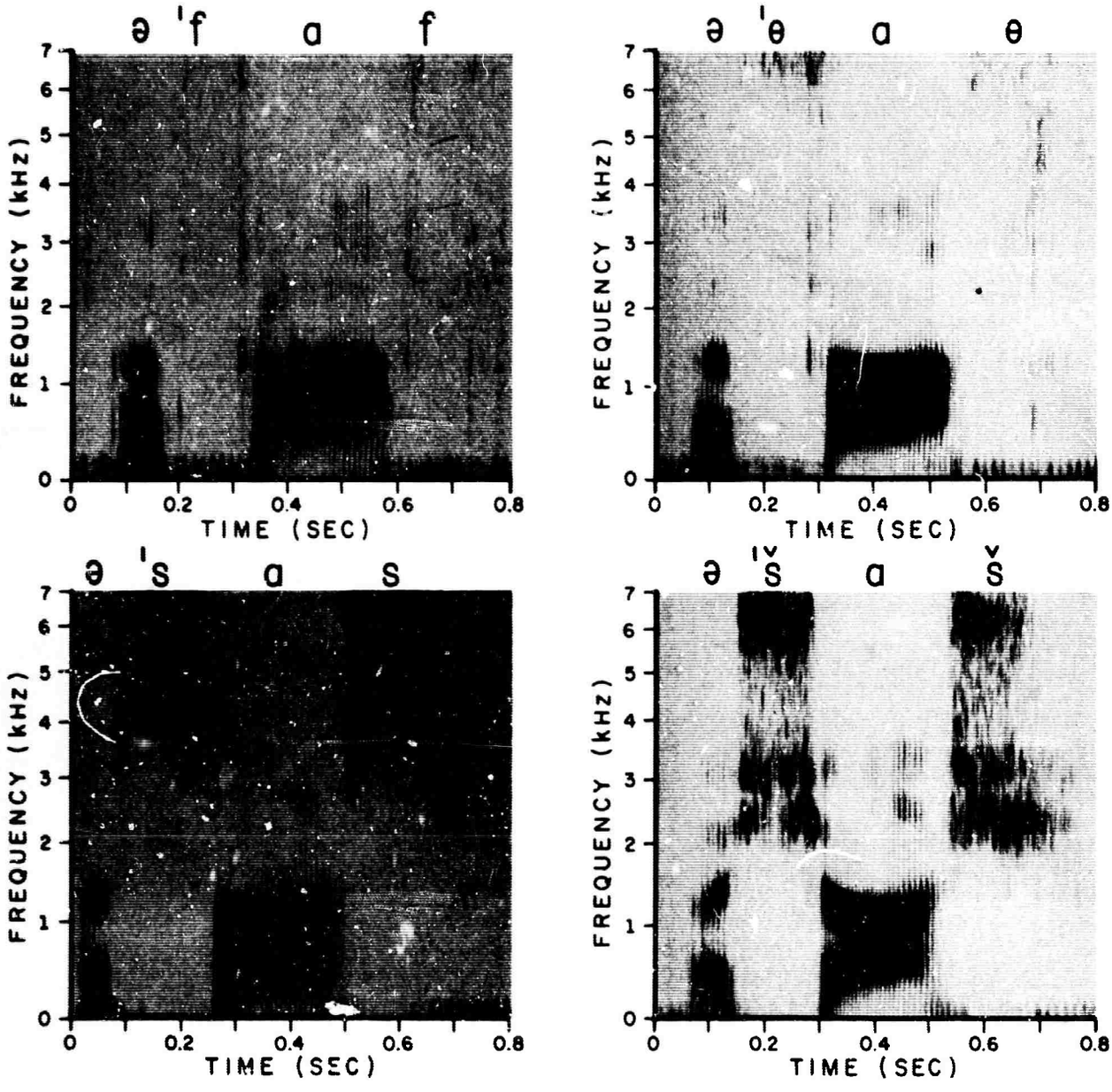


Fig. 11 Spectrograms illustrating properties of voiceless fricative consonants. Speaker KS.

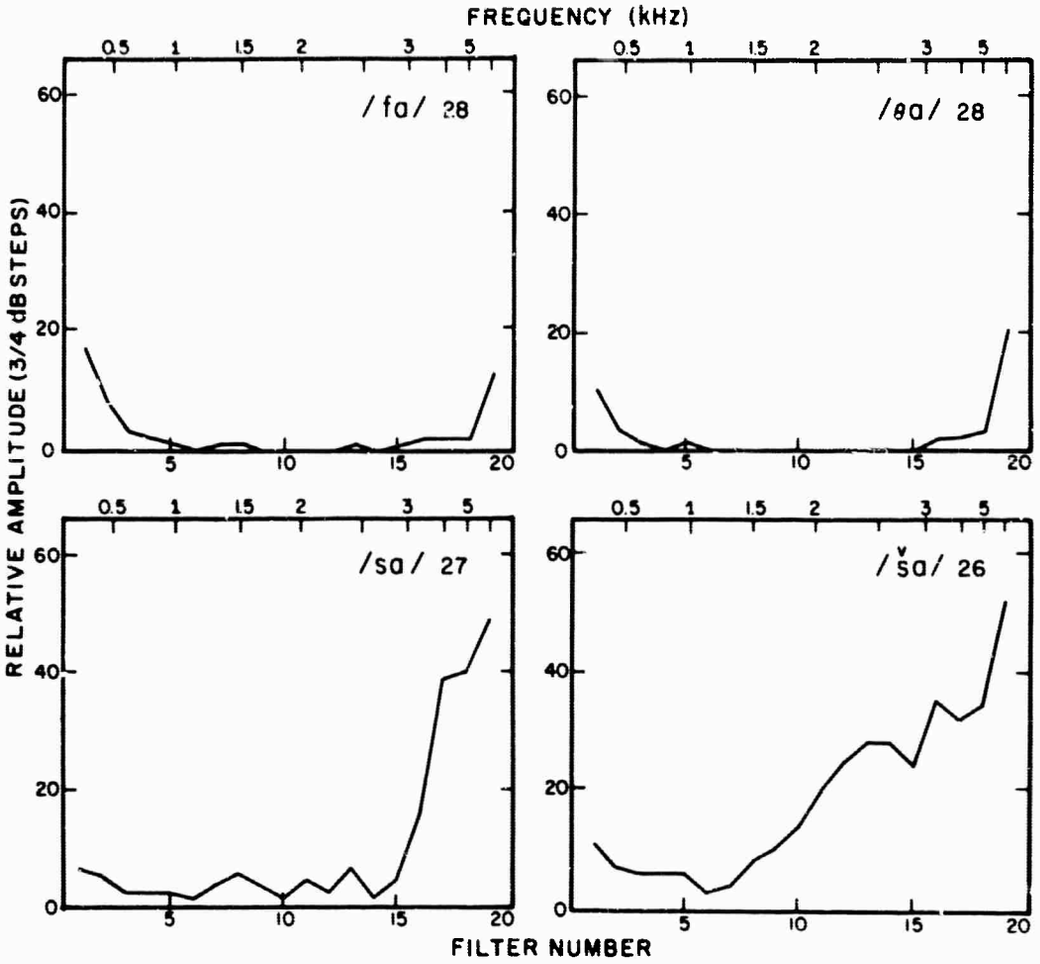


Fig. 12 Spectra during constricted interval for voiceless fricative consonants in pre-stressed position. Speaker KS.

(around 6000 Hz), but the overall intensity for these consonants is quite low. The fricative /f/ has some weak energy in the low-frequency range, but there is essentially no low-frequency energy for the remaining voiceless fricative consonants; this fact provides a simple and reliable way of separating voiceless fricatives from all voiced sounds.

At high frequencies, the spectra for voiced fricatives are very similar to the spectra of their voiceless cognates. The voicing is manifested by low-frequency energy; the amplitude of the first filter output seems to be always the greatest, and there is relatively small output for all low-frequency filters above the second. Examples of spectrograms and spectra for voiced fricatives are shown in Figs. 13 and 14, respectively. The amplitude of the first filter output is consistently 15–30 units below the peak amplitude of the following vowel, depending upon the speaker to some extent. One of the three speakers examined in this study (CW) tends to generate many continuant consonants with a less constricted vocal tract, and consequently the high frequencies (above, say, 500 Hz) for voiced fricatives are not as weak as for the other speakers. An example of a /v/ spectrum for this speaker is shown in Fig. 14.

Durations of the constricted intervals for voiceless and voiced fricatives in prestressed position are somewhat longer than the durations of the closure intervals for the corresponding stop consonants. Examples of these durations obtained from several vowel environments are given in Table VIII. The durations of voiced fricatives are consistently about 50 msec less than those of voiceless fricatives, but the effects on duration of place of articulation are relatively small.

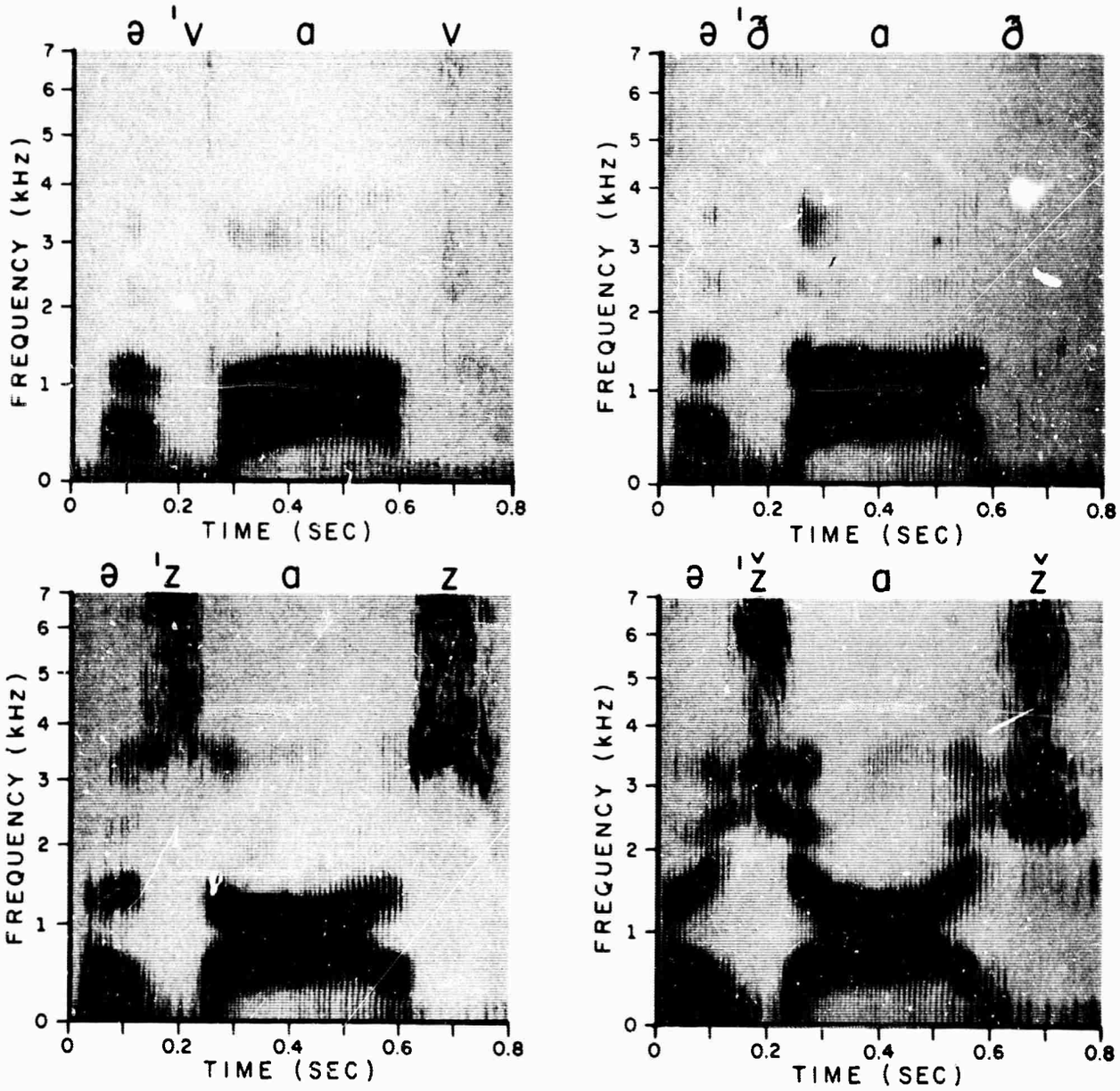


Fig. 13 Spectrograms illustrating properties of voiced fricative consonants. Speaker KS.

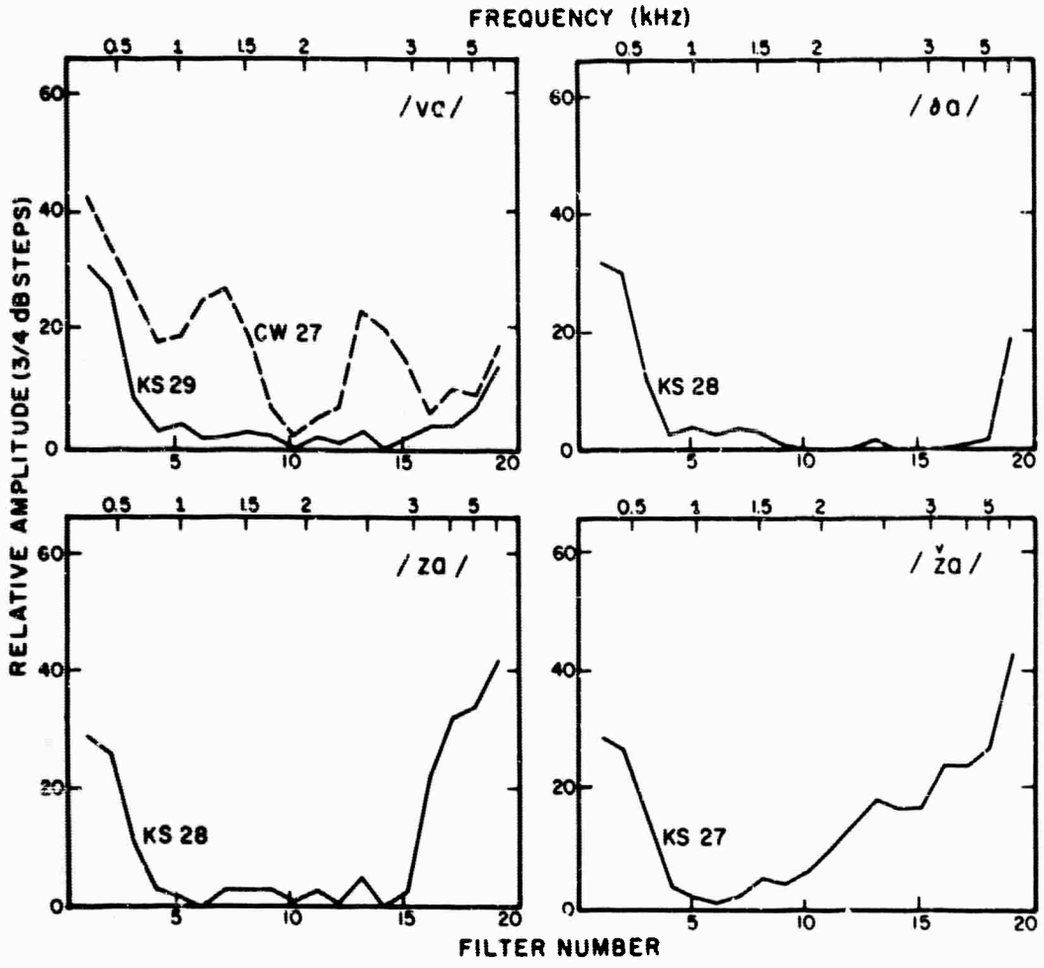


Fig. 14 Spectra during constricted interval for voiced fricative consonants in prestressed position. Speaker KS (except for one spectrum of /v/ as indicated).

TABLE VIII. Durations of constricted intervals for fricative consonants preceding stressed vowels. Averages over three vowel environments /i ɑ u/ and over three speakers.

Consonant	Average Duration (msec)
f	180
θ	180
s	180
s̬	200
v	130
ð	130
z	140
z̬	150

5.3 Constricted interval: liquids and glides (sonorant, non-nasal consonants)

The consonants /w j r l/ in initial prestressed position are all characterized by an interval of 50–100 msec in which the first-formant frequency (frequency of low energy peak) is low and in which the amplitude at low frequencies is lower than in the following vowel. Spectrograms of utterances containing these sonorant consonants are shown in Fig. 15. Examples of spectra taken in the middle of this constricted interval (between initial schwa and stressed vowel in utterances of the type /ə'CV(C)/ are given

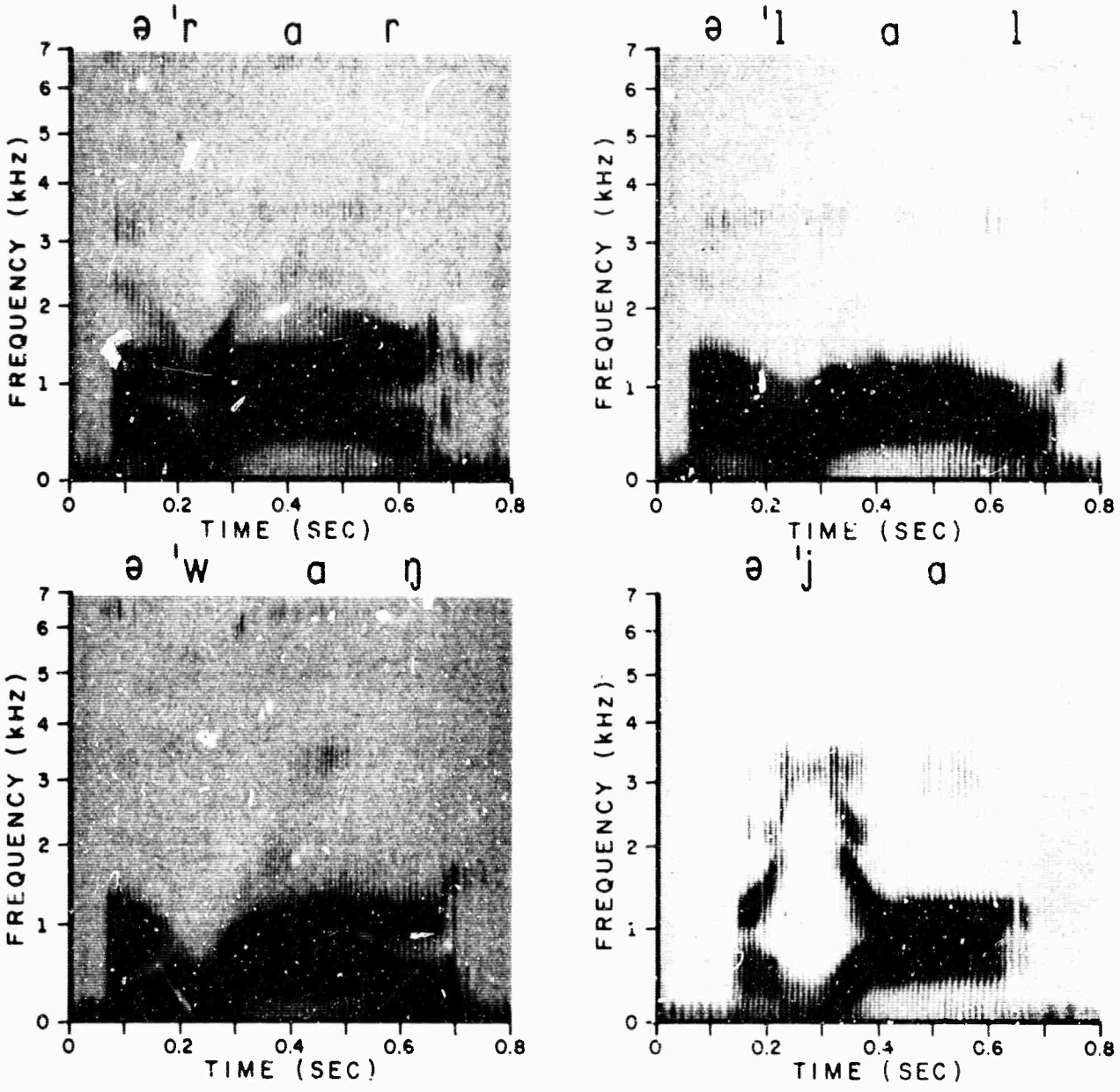


Fig. 15 Spectrograms illustrating properties of liquids and glides. Speaker KS.

in Fig. 16. All of these spectra have a frequency peak either in filter 1 or in filter 2, with sharply dropping energy above this spectral peak. The rate of decrease in energy for filters 3 through 6 is greatest for the glide /j/, and least for the liquids /r/ and /l/. The consonants /r/ and /l/ have two formants in the frequency range up to 1100 Hz, and consequently the low-frequency peak, which is a consequence of the first two formants, is broader than for the /j/, which has only one formant at low frequencies. For the glide /w/, the second formant is lower (it is at about 700 Hz) than for /r/ and /l/, and consequently the drop in energy at the upper side of the low peak is more rapid. For two of the three speakers studied, the initial consonant /r/ seems to exhibit a small secondary peak at filter 5 or 6 (980–1160 Hz), probably because the low third formant helps to boost the amplitude of the second-formant peak.

The peak amplitude at low frequencies for all of these sounds is 5–15 units (4–12 dB) below the peak amplitude of the adjacent vowel. This reduction in energy during the consonant is a consequence of the constricted vocal-tract configuration associated with the consonant; this constricted configuration gives rise to a low-frequency first formant, and it can be shown that the amplitude of the first-formant peak goes down as the frequency of the first formant decreases (Fant, 1956). There may be also some reduction in the output of the vocal cords in the consonantal interval.

Both /r/ and /w/ have essentially no energy in the frequency range above filter 10 (1880 Hz), whereas for /l/ and /j/ there are energy peaks at high frequencies. (For one of the speakers, there is more high-frequency energy in these consonants than for the other

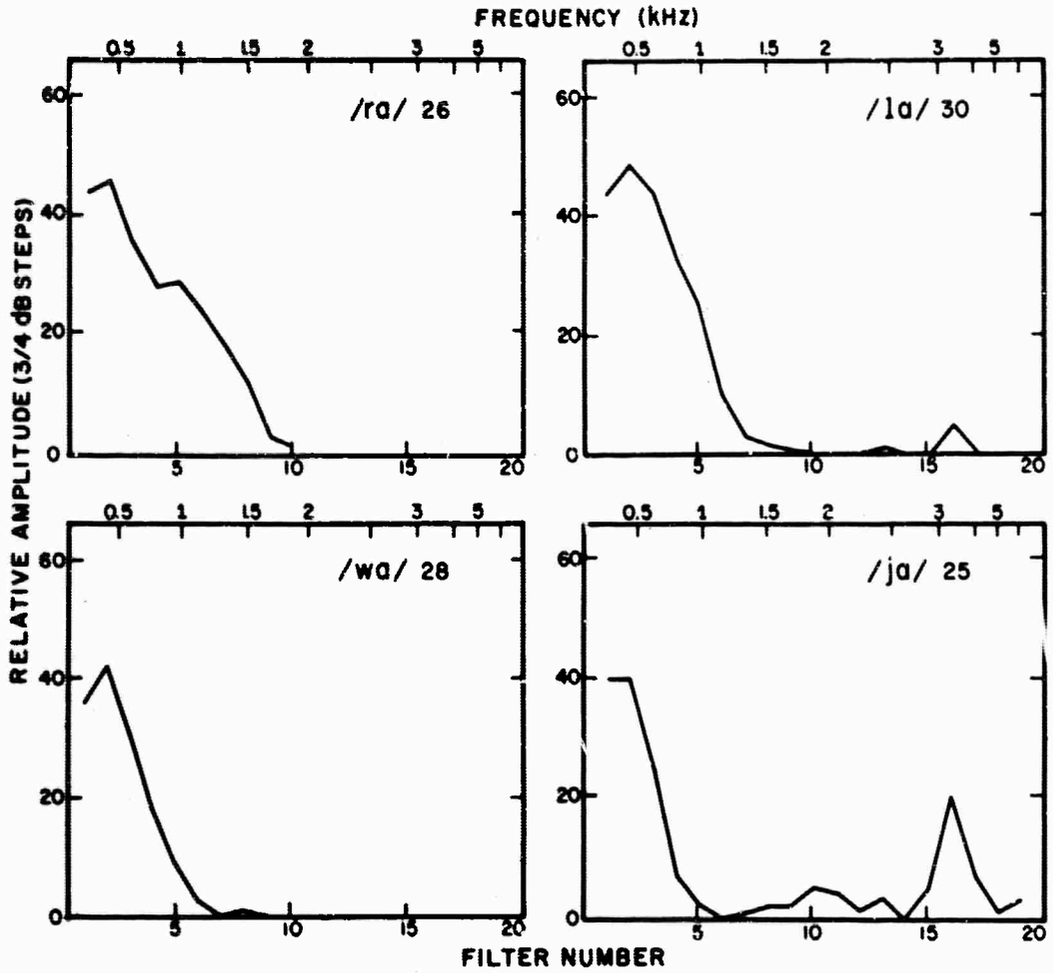


Fig. 16 Spectra sampled during middle of constricted interval for liquids and glides in the pre-stressed position. Speaker KS.

two speakers.) These peaks are more pronounced for the /j/, and the biggest peak seems to be around filter 16 (about 3000 Hz). For /l/ the weak high-frequency peak is in the vicinity of filter 13 (about 2400 Hz). The absolute level of high-frequency energy probably does not provide an important cue for distinguishing among these consonants. Of greater importance, perhaps, is the transition between the consonant and the following vowel, and in particular the contrast in high-frequency energy level between the consonant and the vowel, as discussed later in this Section of the report.

5.4 Release and transitions: stop and nasal consonants

The release from a stop or nasal consonant into the following stressed vowel is characterized by discontinuities in the amplitudes in some frequency regions, as the spectrograms in Fig. 8 have demonstrated, and by transitions of the formants into steady-state positions characteristic of the following vowel. The spectrograms of Fig. 8 show, for example, that the second formant in the syllable /ba/ undergoes a rising transition following release of the consonant, whereas for /da/ and /ga/ there is a falling second-formant transition. At the output of the 19-channel analyzer, the amplitude discontinuities are less obvious that on the spectrograms, since the smoothing filters in the analyzer have relatively long time constants (10-20 msec). The discontinuous changes are expected to occur over an interval as short as 10-20 msec.

Although a criterion for identifying a consonant as being a stop or a nasal has not been worked out in quantitative terms, it is possible to state the nature of such a criterion. The requirement (in the case of a consonant in prestressed position) is that there should be an interval of the order of 50–100 msec or more in duration in which there are no rapid changes in any of the spectrum channels. Following this interval there should be a discontinuous change in most of the spectrum channels in which measurable energy exists. Thus, for example, the amplitude in channel 5 for the utterance /ə'mɑ/, shown in Fig. 17, would satisfy the requirement, whereas a contour like that shown for the utterance /ə'wɑ/ would not satisfy the requirement. Even though the rate of change of amplitude is comparable in the two cases, the glide /w/ does not have a sufficiently long steady-state interval preceding this change. A criterion such as this would, incidentally, place the liquid /l/ in the same class as stops and nasals—a categorization that is appropriate on other grounds.

Nasal consonants can, of course, usually be distinguished from stop consonants by the nature of the spectrum during the closure interval, as noted earlier, although the characteristics of the release into the following vowel also provide important cues for identifying nasals as opposed to voiced stops. Voiced and voiceless stop consonants can often be differentiated on the basis of the very low-frequency energy that may exist during the closure interval for voiced stops but not for voiceless stops. This is not a reliable indicator, however, since frequently a voice bar does not exist for a voiced stop, particularly when it occurs in initial position.

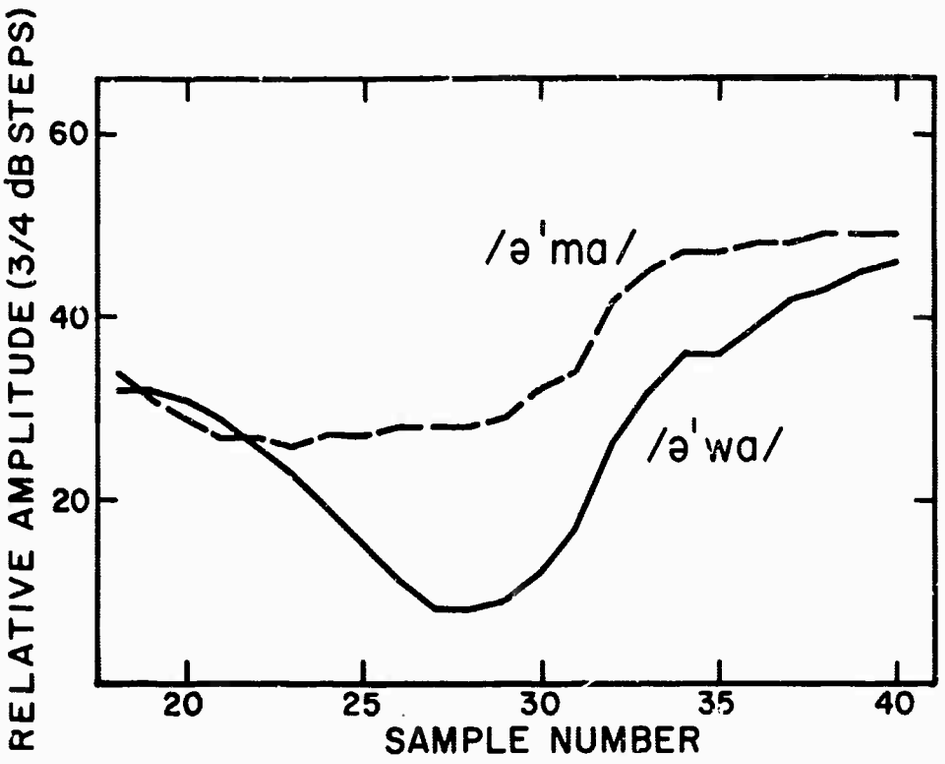


Fig. 17 Graph of amplitude of channel 5 *vs* time for the utterances /ə'mɑ/ and /ə'wɑ/. The jump in amplitude from sample 31 to 32 would be called a "discontinuity" in the case of /ə'mɑ/ because it is preceded by a long interval in which the amplitude is essentially constant

A more reliable procedure for identifying aspirated (voiceless) stops as opposed to voiced stops is to detect the presence of the aspiration noise which always follows the release of a voiceless stop in English. This aspiration interval can be most readily differentiated from voicing by observing the outputs of the lowest three or four filters (region of the first formant of the following

vowel). The amplitude in this frequency range is always less than that for a vowel, by 20 or more units (15 dB or more). This property of voiceless stops is illustrated in Fig. 18, which compares a plot of the output of filter 2 as a function of time for several voiced and voiceless stops. The discontinuities at the consonantal release and again at the onset of voicing for the voiceless stops are clearly observable. The duration of aspiration following the release of voiceless stops is in the range 50–100 msec (Lisker and Abramson, 1964). The duration of frication noise following the release of the affricate consonants /č/ and /ǰ/ is in this range also. The length of aspiration for the voiceless stops tends to be smallest for /p/ and greatest for /k/, although these differences are not observable in the few examples shown in Fig. 18. In the case of the voiced stops, the increase in level immediately following the release is quite rapid since voicing commences immediately upon release or shortly thereafter. This rate of increase in level at low frequencies appears to be greater for the labial stop /b/ than for /d/ or /g/ (as discussed later).

Place of articulation for stop and nasal consonants is determined primarily by the detailed acoustic events at the consonantal release and during the 50- to 100-msec interval following the release. Differences in these transitions between consonant and vowel are attributable, of course, to the fact that the articulatory mechanism must perform movements between the different consonant configurations and the following vowel configurations, and the durations of these movements may be 50 msec or more. If the consonant starting point is different (as it is with the labials /b p m/ as opposed to the dentals /d t n/ as opposed to the velars /g k/) then the transitions will be different.

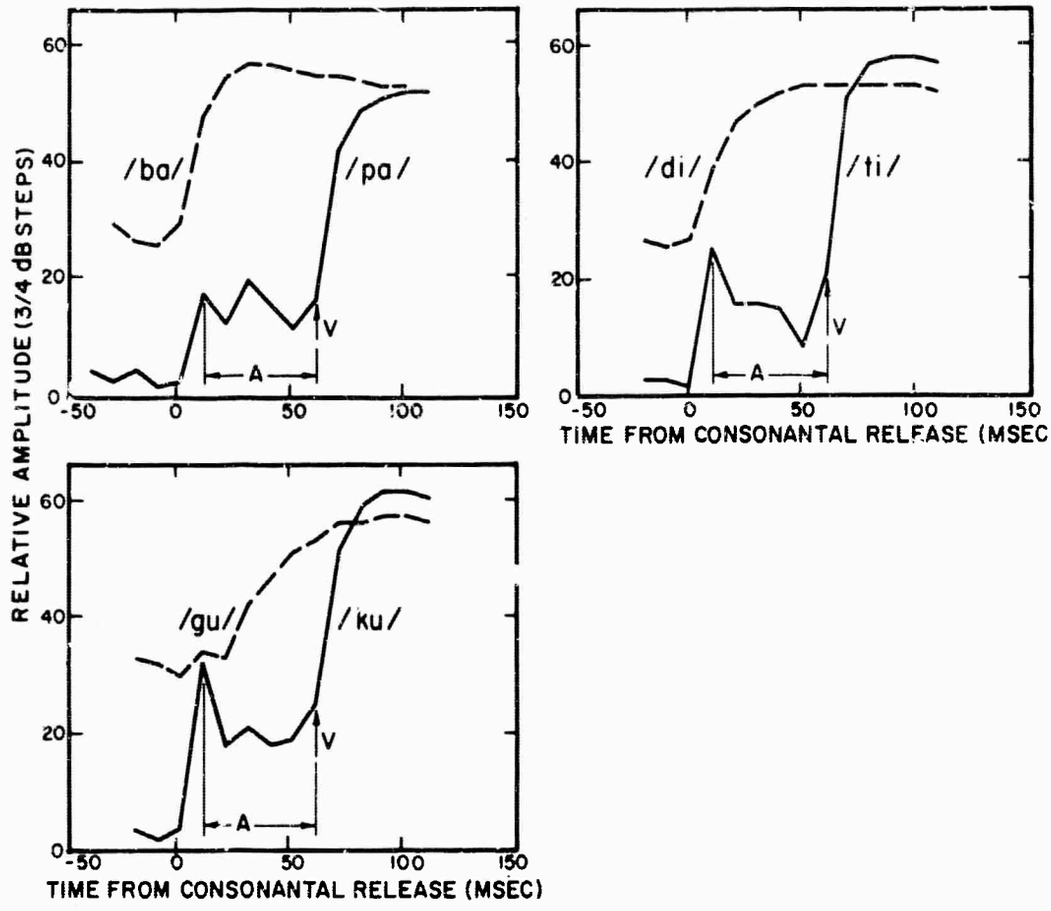


Fig. 18 Output of channel 2 of spectrum analyzer as a function of time during the release of three cognate pairs of voiced and voiceless stop consonants. Speaker KS. The aspiration intervals are indicated by A and the voicing onset times by V.

A measurement procedure that will reliably separate labials, dentals, and velars has not yet been developed. It is known that the transitions of the formants between consonant release and vowel, particularly the second-formant transition, provide important cues for consonant identification (Liberman, Delattre, Cooper, and Gerstman, 1954). The spectral and temporal characteristics of the burst of noise that is often generated at the constriction at the instant of release also are important indicators of place of articulation for the consonant. The spectrograms of Fig. 8 illustrate some of these differences in formant transitions and in the burst, but the differences are subtle and may be difficult to detect by machine. Furthermore, the directions of the formant transitions for a given consonant may depend to some extent on the following vowel.

An illustration of the kind of procedure that will be necessary to identify these consonants is presented in Fig. 19. This Figure shows the outputs of three of the filters (filter 5 at 980 Hz, 8 at 1520 Hz, and 16 at 3260 Hz) as a function of time in the 100-msec interval immediately following the release, for the syllables /ba/, /da/, /ga/, /na/, and /ma/. For /ba/, there is a tendency for the amplitude increases in filters 8 and 16 to lag behind that in filter 5. In the case of /da/, on the other hand, filter 16 shows the earliest initial onset, with the amplitudes in filters 8 and 5 rising at successively later times. Filter 8 shows the most rapid initial rise for the syllable /ga/, and this rate of increase is more abrupt than for /ba/ and /da/. Thus, there is a tendency for the initial onset of energy to be at low frequencies for /b/, at high frequencies for /d/, and in the midfrequency range for /g/ (Stevens, 1967). The initial /n/ in /na/ shows some of the characteristics of /d/: there is negligible energy in

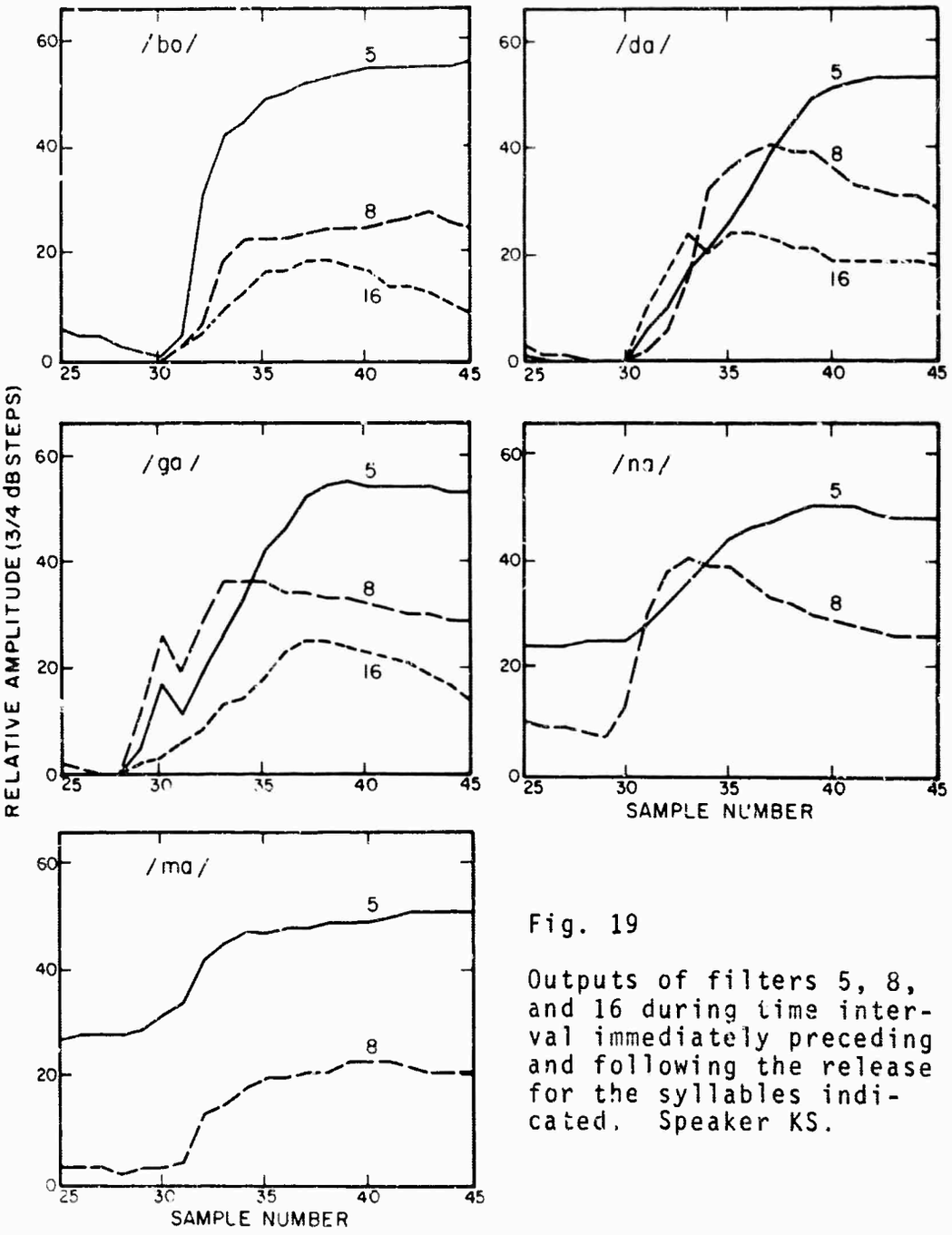


Fig. 19

Outputs of filters 5, 8, and 16 during time interval immediately preceding and following the release for the syllables indicated. Speaker KS.

filter 16 in this case, but the onset of energy in filter 8 precedes that in filter 5. This is not the case for the syllable /mɑ/. For the /g/ in prestressed position, there is an initial burst of noise energy of about 30-msec duration; evidence of this noise burst can be seen in channels 5 and 8 of Fig. 19. The initial burst is briefer for /d/ (about 10 msec) and of higher frequency (as seen in channel 16 of Fig. 19), but there is essentially no noise burst at the onset of /b/. These properties of the burst can also be observed qualitatively in the spectrograms of Fig. 8. The picture presented in Fig. 19 will change somewhat depending upon the following vowel, and would probably show the effects more clearly if the averaging times of the smoothing filters in the analyzer were shorter.

For initial voiceless stop consonants, the time-varying spectral patterns immediately following the release are similar in some respects to those of the voiced stops. Figure 20 shows spectra sampled at 60-msec intervals following release of each of the voiceless stops with the following vowel /ɑ/. In the case of the labial stop /p/, the spectrum in the aspiration interval immediately following the release indicates weak energy with no pronounced high-intensity spectral peaks. After onset of voicing, the vowel spectrum remains reasonably stable. For /t/ and /k/, on the other hand, there are pronounced spectral peaks in the noise interval, at high frequencies for /t/ and in the midfrequency range for /k/. The spectrum changes that occur after voicing onset indicate that there is a rising transition of the first formant and a falling transition of the second formant during the /ɑ/ following these two consonants. The spectrum characteristics during the aspiration interval for /k/ are very much dependent upon the following vowel.

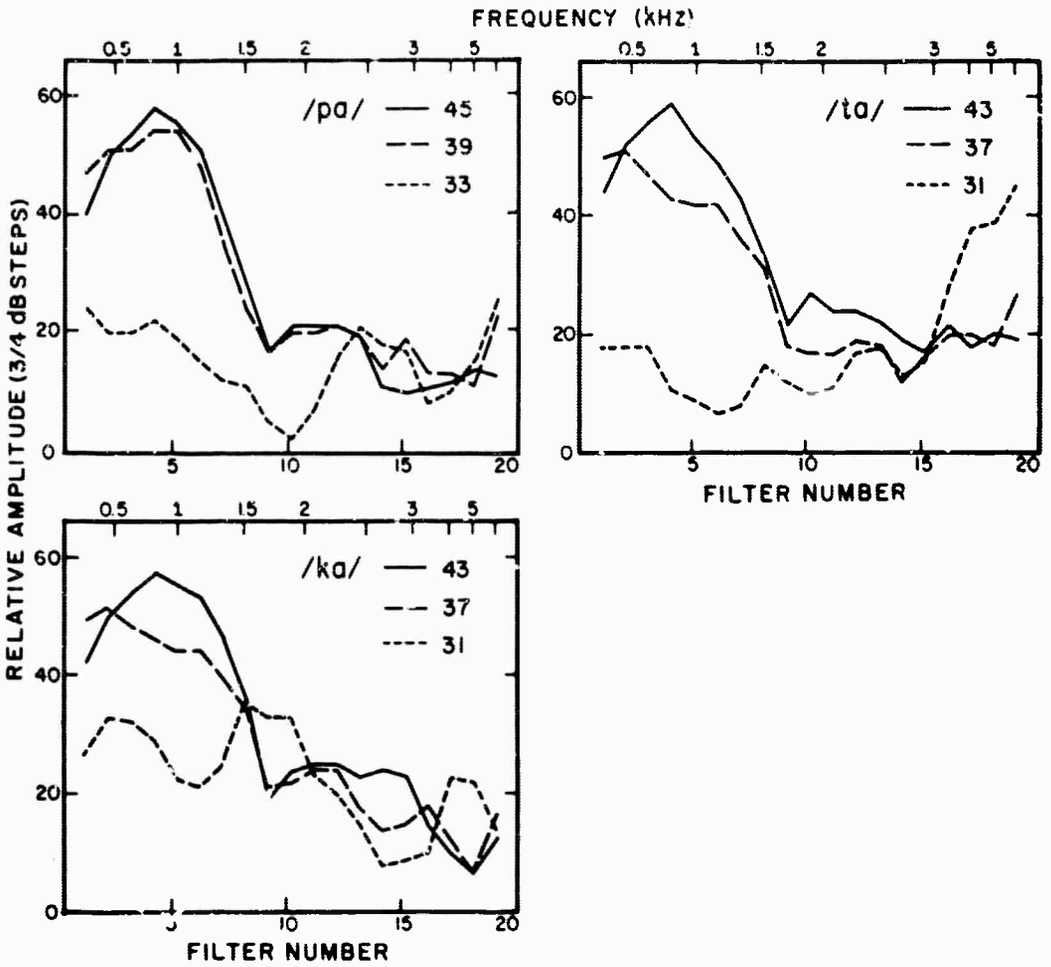


Fig. 20 Spectra (obtained from the 19-channel filter bank) sampled at 60-msec intervals following release of the stop consonants in the syllables /pɑ/, /tɑ/, and /kɑ/. The first spectrum is sampled about 20 msec after consonant release, the second spectrum occurs about 20 msec after voicing onset and the third is in the middle of the vowel. Speaker KS.

5.5 Release and transitions: liquids and glides

As noted above, the consonants /w j r/ in the prestressed position do not exhibit a steady-state interval during the constriction, but rather are characterized by continuous change. This property is illustrated in the spectrograms of the consonants shown in Fig. 15. Examples of several filter outputs, plotted as a function of time, for /w j r l/ in the environment /ə'Cɑ/ are displayed in Fig. 21. For /wɑ/ the amplitude rise for filter 8 occurs later than for filter 5 (as it does in the syllable /bɑ/), whereas for /jɑ/ the situation is reversed. As noted earlier, there is a reasonably long steady-state interval for /l/ before the rapid transition into the following vowel. No such steady-state region exists for /r/, but with /r/ there is a greatly delayed onset of the amplitude of filter 16 relative to lower-frequency filters (as a consequence of the rising transition of the third formant, which is always characteristic of initial /r/).

When the consonants /w j r l/ occur in other vowel environments, data similar to those shown in Fig. 21 are obtained, but the relations between timing of filter outputs may not always be as clear as those in the Figure.

5.6 Summary of characteristics of consonants in prestressed position

The acoustic attributes of consonants in prestressed position can be conveniently summarized in terms of the features listed in Table VI.

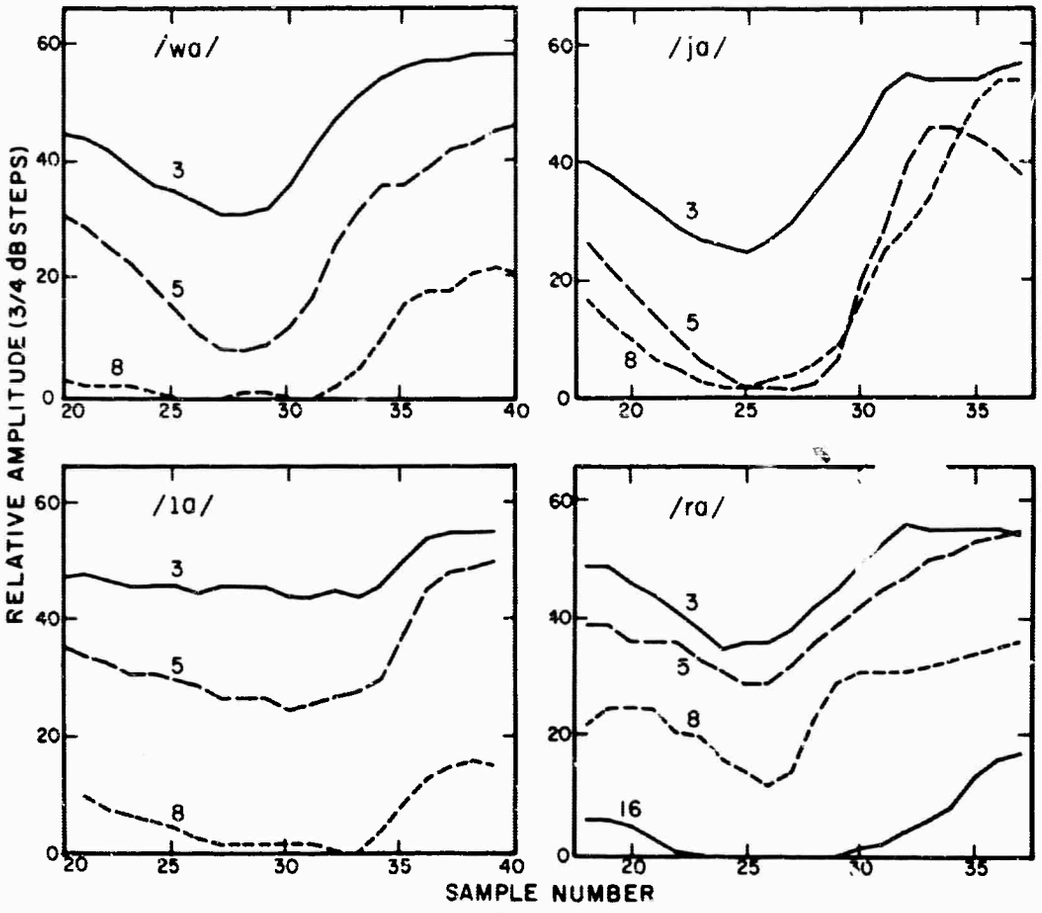


Fig. 21 Outputs of several filters (as indicated) for liquid and glide consonant in the environment /ə'Cɑ/. Speaker KS (in all cases).

Stop consonants are characterized by a closure interval within which the properties of the signal at low frequencies do not change appreciably and in which there is negligible high-frequency energy (above, say, 2000 Hz). In the case of nonnasal stops, there is essentially no sound energy in this interval above about 500 Hz. This closure time is followed by an almost discontinuous change coincident with release of the articulatory closure into the following vowel. This abrupt change occurs in some frequency ranges, but not necessarily at all frequencies, particularly in the case of the sonorant stops (i.e., nasals and /l/).

For *sonorant* consonants, voicing continues with appreciable energy through the closure interval, and the spectral maximum during this interval is always in the first or second filter of the 19-channel analyzer used in this study, i.e., below 440 Hz. The spectral energy in this frequency range is several decibels below that of the following vowel. A sonorant has no high-frequency noise energy.

The feature *voicing* implies that periodicities and low-frequency energy continue through the closure interval. In the case of voiced stop consonants, vocal-cord vibration may be weak or absent during the closure interval, but voicing commences almost immediately upon release of the stop. Thus, for a segment to be voiced, an interval of high-frequency noise (30 msec or more in duration) must not occur in the absence of low-frequency periodicities.

As Table VI indicates, a fricative consonant has the features *-stop, -sonorant*. Thus, such a consonant has high-frequency noise energy during the closure interval, and may or may not be voiced during this time.

The acoustic correlate of the feature *aspiration* (which in English applies only to stop consonants) is an interval of 50–100 msec of noise that occurs after release of the stop and before voicing occurs. This aspiration noise has negligible energy in the low-frequency (first-formant) region.

A *nasal* consonant is a sonorant and a stop consonant, and hence has the attributes of both of these features. A distinguishing attribute of a nasal is a relative lack of energy in the frequency region around 800 Hz, whereas there is strong spectral energy at lower frequencies and there may be spectral peaks above 800 Hz.

The features of place of articulation for consonants (the features *anterior* and *coronal* in Table VI) are difficult to describe simply in terms of common properties that are valid for fricative, stop, and sonorant consonants. For fricative consonants, the noise spectrum during the constricted interval provides important cues for place of articulation. For other classes of consonants, place of articulation is determined by the way in which the characteristics of the signal change with time in the few tens of milliseconds following consonant release. Postdental consonants tend to have strong initial high-frequency energy in this interval, whereas for velars the energy onset is in the midfrequency range. Further study is needed to provide a more precise specification of the acoustic correlates of place of articulation, particularly for stop consonants.

6. CONSONANT CLUSTERS IN PRESTRESSED POSITION

The consonant clusters that can occur in initial position in English can be divided into two classes: (1) /s/ followed by a stop (or nasal) consonant (/sp/, /sn/, etc.); and (2) a liquid or glide preceded by a stop or fricative consonant (/pr/, /fl/, /gr/, /tw/, etc.). A third class consists of triplets that are members of both these classes, e.g., /str/, /spl/. Examples of spectrograms of these clusters in the environment /ə'$C_1C_2(C_3)$ɑ/ are shown in Fig. 22.

In the case of clusters with initial /s/, the fricative has more or less the same spectral characteristics as an initial fricative without an adjacent consonant. Its duration tends to be somewhat shorter, however, particularly when it precedes a stop consonant. Measured durations of the noise in /s/ in such clusters are given in Table IX. The noise interval appears to be the shortest for the three-segment clusters and longest preceding /l/ and /w/. A stop consonant following /s/ has very little aspiration noise following the release. Values of the duration of the noise interval between release of the stop and the onset of voicing are about 30 msec for velar consonants and less for other stops. The durations of the stop gap and of the nasal murmur (for the clusters /sm/ and /sn/) are also given in Table IX. These durations are considerably shorter than the corresponding durations when the stops and nasals occur as single prestressed consonants. The discontinuous changes at the release of the stop and nasal consonants are similar to those discussed previously.

Table IX also shows the approximate durations of the constricted interval in sonorant consonants when they are preceded by /s/.

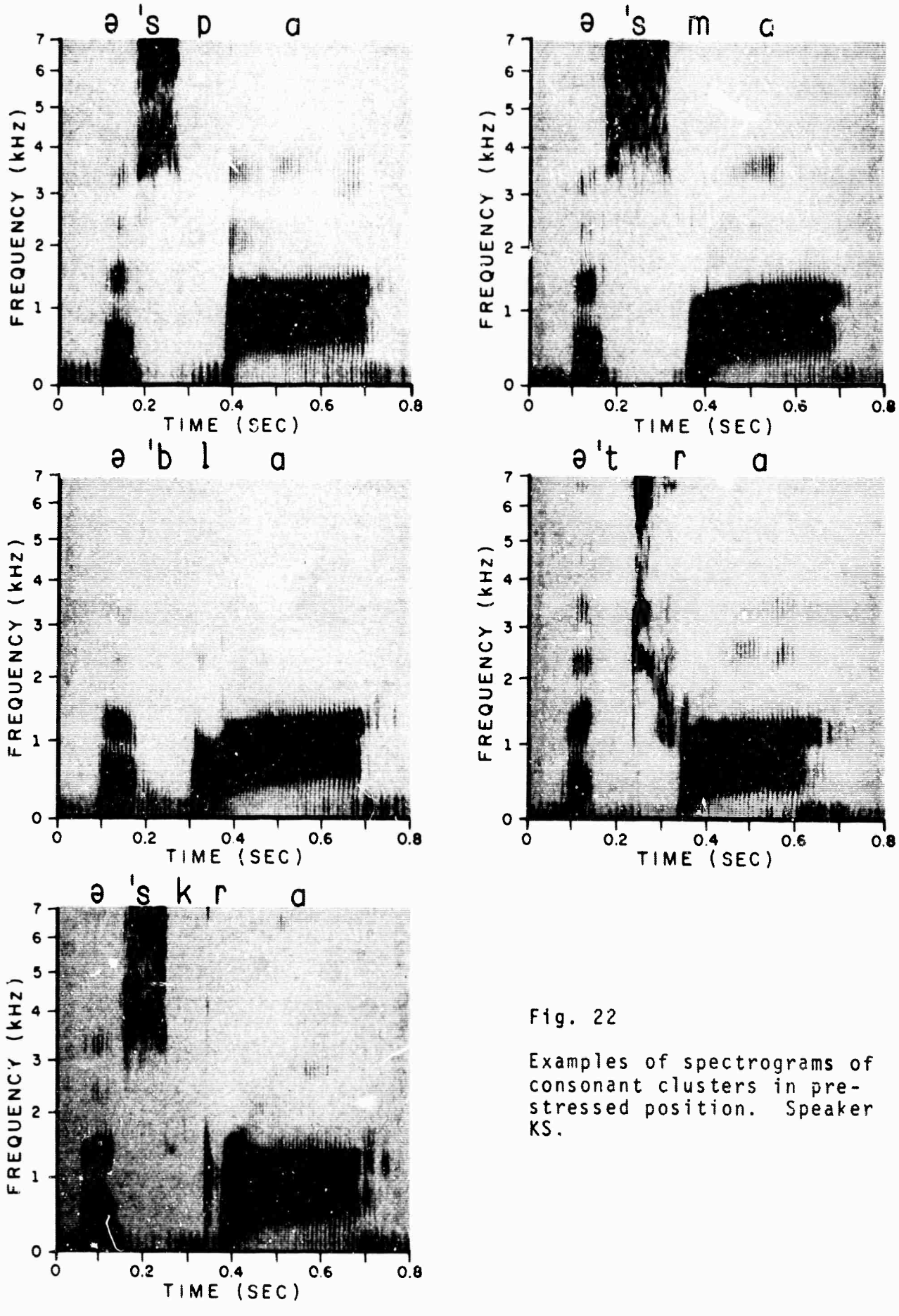


Fig. 22

Examples of spectrograms of consonant clusters in pre-stressed position. Speaker KS.

Again, this duration (average of 50 msec) is shorter than the corresponding duration when the sonorant is the only consonant preceding a stressed vowel.

TABLE IX. Durations of noise in fricative /s/ and of stop gap or sonorant murmur in following consonant for various consonant clusters in the environment /ə'sC(C)ɑ/. Average values for three speakers.

Cluster	Duration of Noise (msec)	Duration of Stop Gap or Sonorant Murmur (msec)
sp, st, sk	120	80
sm, sn	150	50
sl, sw	160	~50
skr, spl	110	90 (stop gap)

When a voiceless stop consonant is followed by a glide or liquid, the duration of the closure interval is slightly less than that for a stop immediately preceding a vowel but the duration of the aspiration is considerably greater. The durations of aspiration measured from spectrograms of the utterances containing such clusters range from 80 to 110 msec, compared with 50 to 100 msec when the consonants appear singly. The duration of the voiced segment of the glide or liquid preceding the vowel is quite brief. After the onset of voicing, a rapid transition toward the following vowel begins almost immediately. These characteristics are observable in the spectrograms of Fig. 22.

7. UNSTRESSED VOWELS AND VOWELS WITH SECONDARY STRESS

A vowel that forms the nucleus of a syllable in English can be assigned a feature which indicates the stress of the syllable. It is possible for two utterances to be identical in all respects except the stress on the syllabic nuclei. For example, the noun and verb forms of the word *reject* differ only in the stress assigned to the two syllables.

Listeners appear to be able to assign at least three degrees of prominence to vowels that form the nuclei of syllables in English. It is generally assumed, then, that a syllable can be characterized by at least three degrees of stress. The degree of stress on a vowel determines, to some extent, the acoustic characteristics of the vowel, particularly its duration, its fundamental frequency, and its intensity. As has been noted earlier, however, stress on a vowel also has a marked effect on the properties of the consonants that precede and follow the vowel.

The acoustic characteristics of vowels and consonants that have been discussed up to this point were obtained for syllables with primary stress, i.e., with the highest of the three degrees of stress. We use the term *secondary stress* to designate the degree of stress on a vowel whose quality is similar to that of a vowel with primary stress but whose prominence, as judged by listeners, is less than that of some other vowel in the same utterance that is judged to have primary stress. Thus, in each of the words *seaweed, essay,* and *cocoa,* the second vowel is considered to have secondary stress.* A still lower degree of stress is assigned to

*This designation of stress is not entirely in accord with that of others, but is sufficient for our purposes.

a vowel for which the quality is changed to that of a schwa vowel /ə/ by virtue of this stress assignment. The vowel in this case is said to be *reduced*. Thus the second vowel in the word *famous* is generally regarded to be a reduced vowel. For a reduced vowel, it is not necessary to specify place of articulation, since two words cannot differ only in the place of articulation of a reduced vowel.

Other examples of reduced vowels are the vowels in the second syllables of the words *poker*, *bushel*, and *wagon*. The unstressed vowels in these words are often designated as syllabic /ɹ̩/, /l̩/, and /n̩/, respectively, although it may be more appropriate to represent such vowels as a schwa vowel /ə/ followed by a final consonant /r/, /l/, or /n/.

7.1 Duration and fundamental frequency

In general, unstressed vowels tend to be shorter and of lower amplitude than stressed vowels, but the words studied here do not provide enough examples of the vowels to permit quantitative data on these durations and amplitudes to be tabulated. Durations of reduced vowels in syllable-final position in bisyllabic words are in the range 50–150 msec. Vowels with secondary stress tend to be slightly longer. When a reduced vowel occurs in initial position in a bisyllabic utterance (as in the utterances /ə'CVC/ or as in a word like *about* or *alike*), the duration can become as short as 20 msec, and the vowel may consist of just two or three glottal vibrations. In fact, it is not uncommon for such a vowel to be omitted altogether in rapid speech.

Fundamental frequency for vowels in unstressed syllables tends to be lower than for stressed vowels. Although detailed measurements of fundamental frequency were not made in this study, informal observations of spectrograms support this result, which has often been reported in other studies. (See, for example, Lieberman, 1967.) Neither fundamental frequency nor duration provide a reliable procedure, however, for identifying the degree of stress that is assigned to a vowel.

7.2 Spectra of vowels with secondary stress

Examples of spectra for nonreduced vowels generated by one of the speakers are shown in Fig. 23. Comparison of these spectra with those shown earlier for stressed vowels (Fig. 5) indicates that the unstressed vowels have similar frequency characteristics. The peak amplitude for the unstressed vowel tends to be lower than that for the stressed vowel in the same word, but the amplitude differences between stressed and unstressed vowels are by no means consistent for all three speakers. While the spectrum shape is similar for the same vowel with primary and with secondary stress, the spectrum of the latter sometimes has weaker low-frequency energy relative to high-frequency energy. This effect may, however, be due to the fact that the vowel with secondary stress is always in utterance-final position for the words examined here.

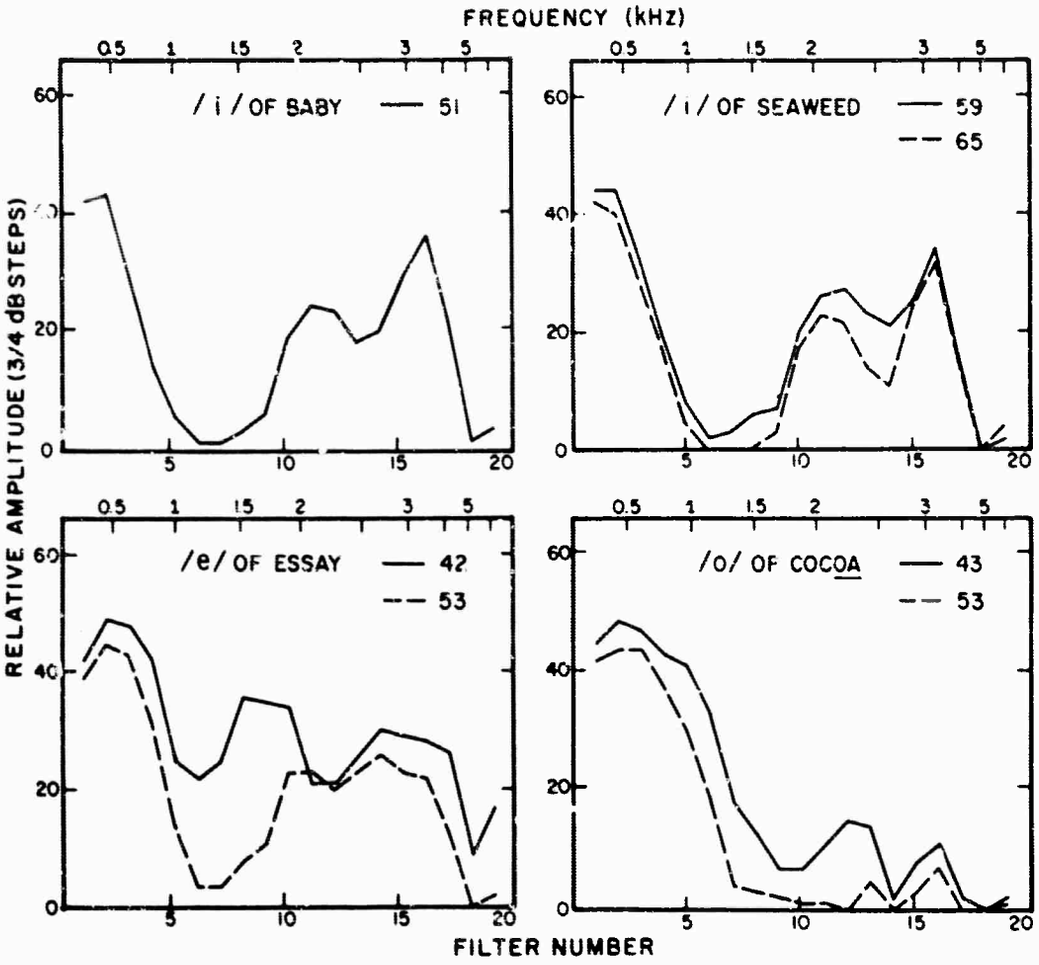


Fig. 23 Spectra of vowels with secondary stress. The solid lines represent spectra sampled in the early part of the vowel; the dashed lines represent spectra sampled later in the vowel.

7.3 Spectra of reduced vowels

Spectra sampled within several reduced vowels of various types are plotted in Fig. 24. All of these vowel spectra are characterized by a low-frequency peak at filter 2 or at filter 3. The schwa vowel in these examples has a second-formant peak at filter 8 (1520), as is characteristic of such neutral vowels, but the position of the second formant for such a vowel may vary appreciably with context. The final reduced vowel (in the word *ragged*) appears to be richer in high-frequency energy than the initial reduced vowel (in the utterance /ə'dad/). Syllabic /l̩ r̩ n̩/ have much less high-frequency energy than the schwa vowel. All of these syllabic sounds appear to have more high-frequency energy than their consonantal counterparts that occur in prestressed position. Presumably the syllabic sounds tend to be generated with a more open vocal-tract configuration, and are therefore characterized by a higher first-formant frequency and possibly by a vocal-cord source spectrum that is richer in high frequencies. The schwa vowel in prestressed position (as in /ə'dad/), on the other hand, is probably generated with a more constricted vocal-tract configuration and with a vocal-cord pulse shape that is broader. Both of these effects would tend to reduce the amount of high-frequency energy relative to the energy at low frequencies.

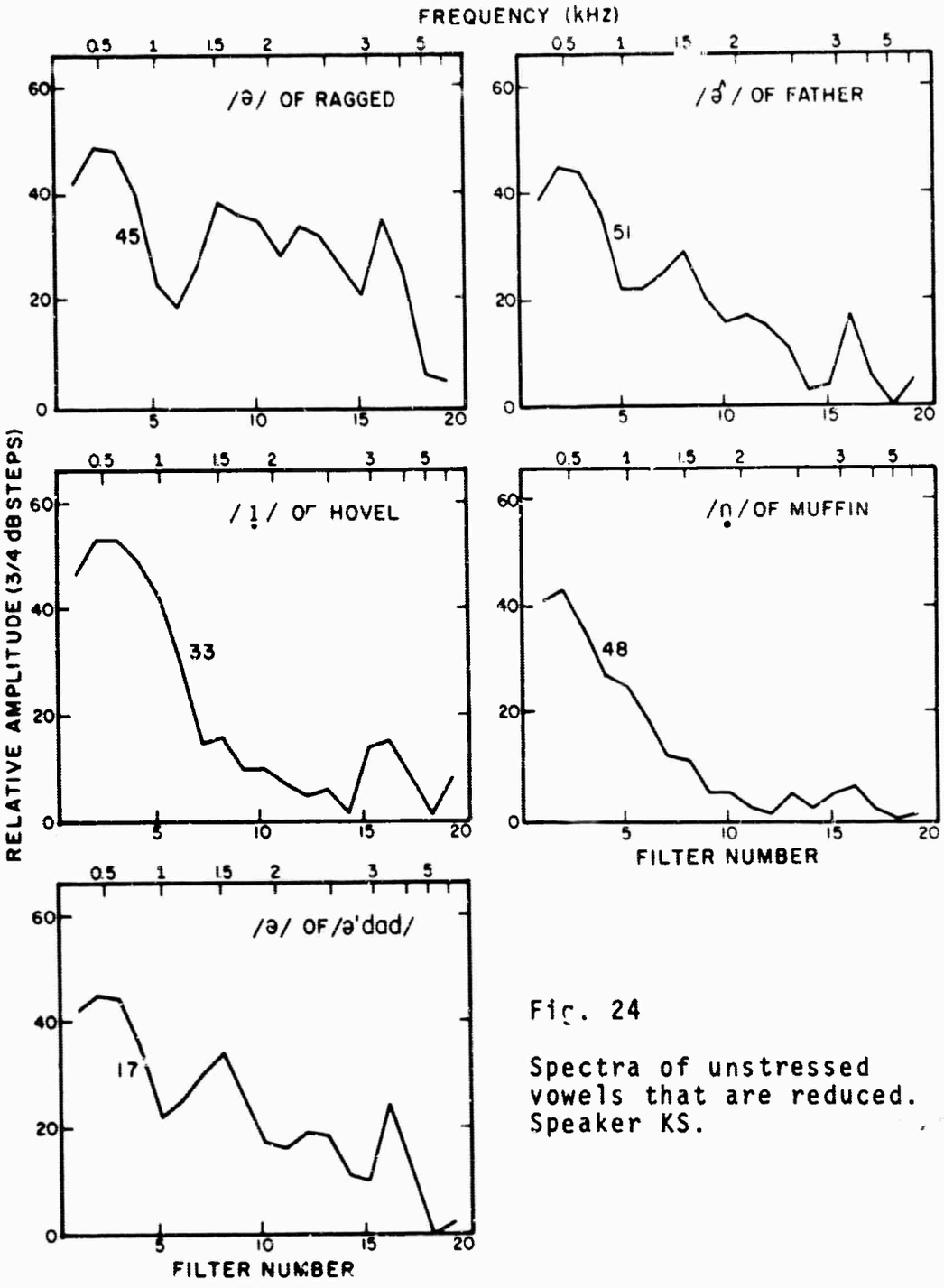


Fig. 24

Spectra of unstressed vowels that are reduced. Speaker KS.

8. CONSONANTS IN POSTSTRESSED POSITIONS

The acoustic characteristics of consonants in poststressed positions can be altered drastically relative to those in prestressed position. A complete discussion of all the ways in which modifications of consonant properties can occur cannot be given in this report, but a few examples can be cited.

When nasal or stop consonants occur after stressed vowels and preceding unstressed vowels, the duration of the closure interval may be greatly reduced. Values in the range 15–110 msec are observed on spectrograms, as opposed to 110–130 msec when the consonants are in prestressed position (in the environment /ə'CV/). Examples of these brief closure intervals are shown in the spectrograms in Fig. 25. These durations are particularly short (15–50 msec) for dental consonants /d t n/ followed by reduced vowels. The duration of aspiration at the release of a stop consonant into an unstressed vowel is also greatly reduced, and in some cases a voiceless stop consonant may have essentially no aspiration in this environment.

Likewise the durations of fricative consonants are reduced in a poststressed environment preceding an unstressed vowel, although the reduction is not as marked as it is with some stop and nasal consonants. Durations of voiceless fricatives are in the range 110–150 msec, as opposed to about 180–200 msec in a prestressed environment; voiced fricative durations are 70–120 msec, in contrast to 130 msec or more in a prestressed environment.

The spectra of consonants in poststressed position preceding unstressed vowels are considerably different from spectra of the

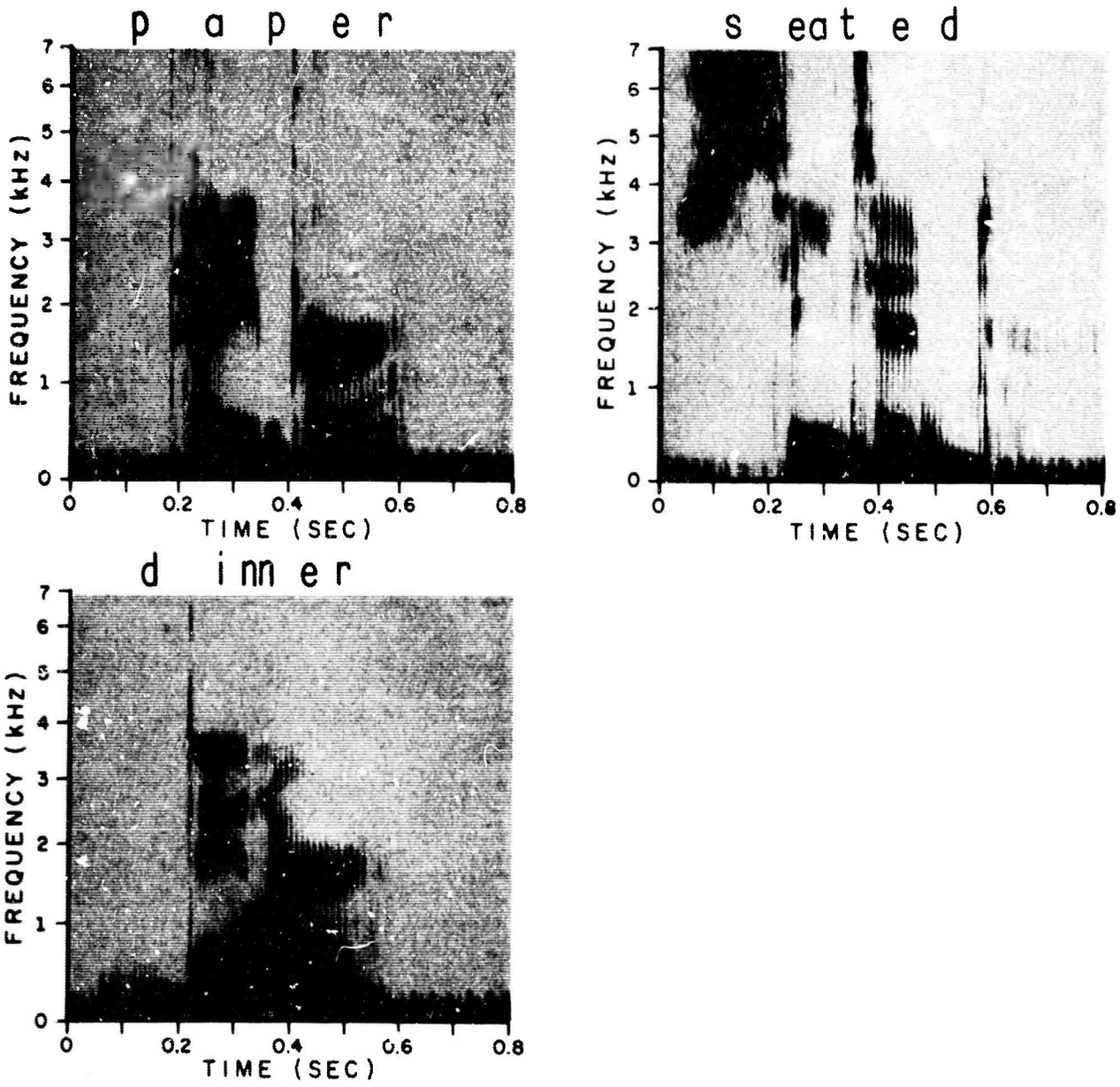


Fig. 25 Spectrograms illustrating acoustic characteristics of consonants in poststressed position preceding reduced vowels. Speaker KS.

corresponding prestressed consonants. Voiced consonants exhibit much more high-frequency energy during the closure interval, and also have a greater overall amplitude. Figure 26, for example, shows spectra sampled during the closure interval for several voiced fricatives, nasals, liquids, and glides. Also plotted on these graphs are examples of spectra for the same consonants in prestressed position (shown previously in Figs. 10, 14, and 16). The differences are presumably due in part to the fact that fricatives, glides, and liquids are not as tightly constricted in poststressed position as in prestressed position, with the result that the low-frequency peak is not as low in frequency. It may also happen that the spectrum of vocal-cord vibration is richer in high-frequency energy for the poststressed consonants. Voiceless stop consonants in poststressed position, on the other hand, are characterized by spectra that are quite similar to those in prestressed position.

For some of the bisyllabic utterances with stop consonants in poststressed position, there is evidence from the spectrograms that the stop closure was not complete. This is particularly true of velar stop consonants (/kg/). In other words, the feature *stop* is often not clearly registered in the acoustic signal for this phonetic environment, except that the closure interval for these stops is generally shorter than that for fricatives.

The acoustic properties of the release of various consonants into unstressed vowels have not been examined in detail in this study. It is evident, however, that the kinds of data that have been shown earlier for prestressed consonants are substantially modified and blurred when the consonants precede unstressed vowels.

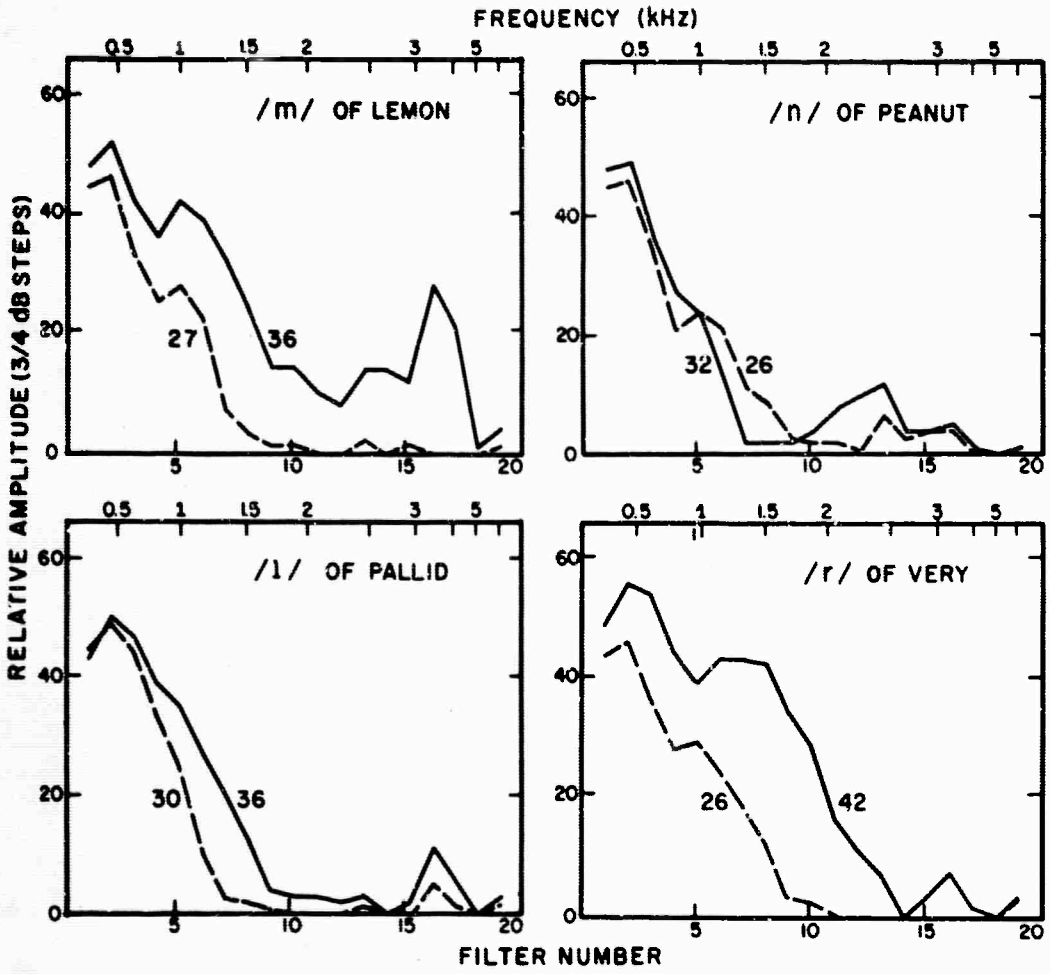


Fig. 26 Spectra of some voiced consonants during the closure interval (solid lines). The consonants are in poststressed position preceding reduced vowels. Shown for comparison are spectra of the same consonants in the prestressed environment /ə'Cɑ/ (dashed lines). Speaker KS.

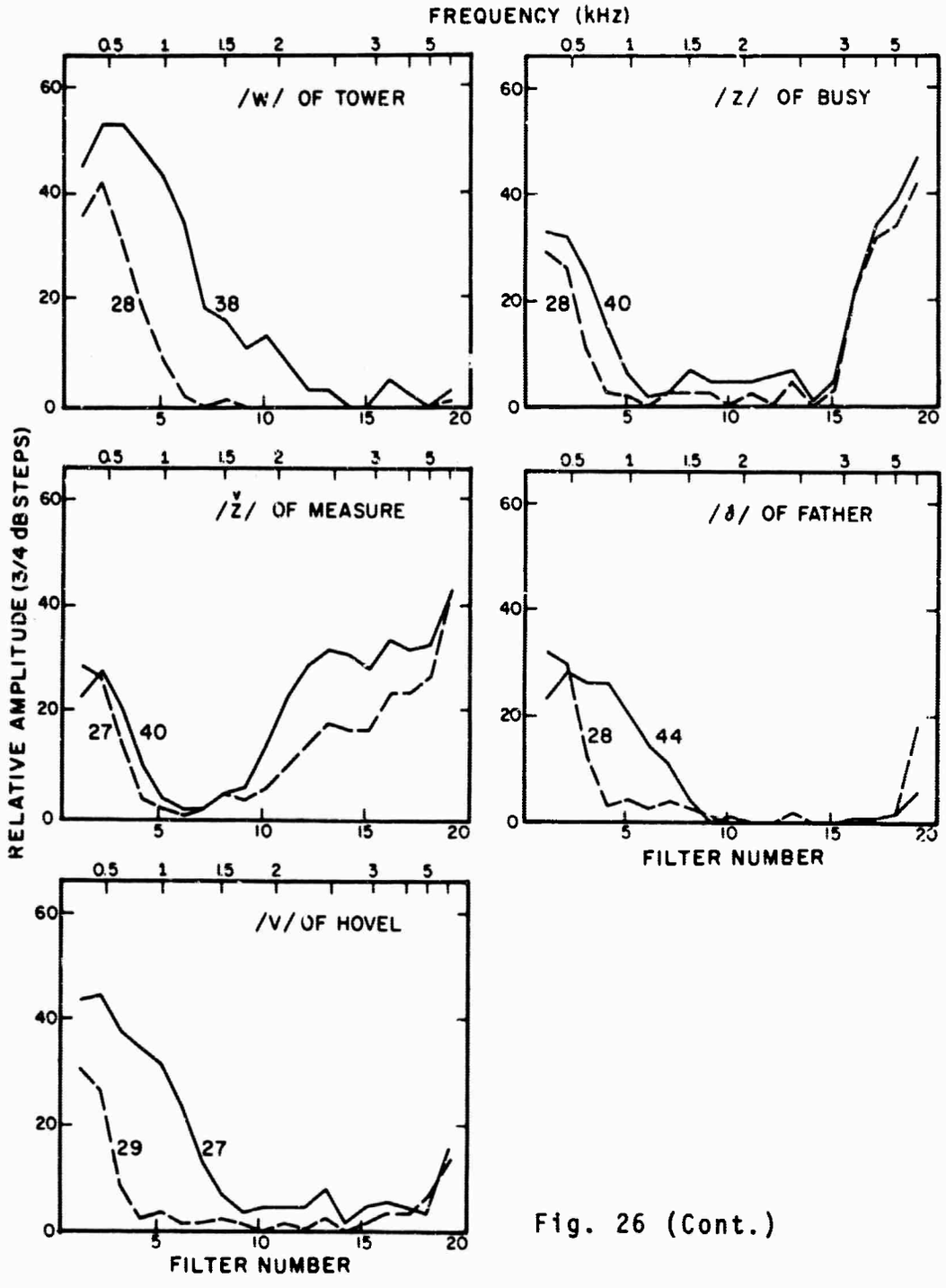

FREQUENCY (kHz)
0.5 1 1.5 2 3 5
/W/ OF TOWER
28
38
/Z/ OF BUSY
28
40
/ž/ OF MEASURE
27
40
/ð/ OF FATHER
28
44
/V/ OF HOVEL
29
27
RELATIVE AMPLITUDE (3/4 dB STEPS)
0 20 40 60
5 10 15 20
FILTER NUMBER

Fig. 26 (Cont.)

The speech material used in this study provides many examples of consonants in word-final position. The durations of these consonants, which occur at the ends of breath groups in all of the utterances, are subject to considerable variability. Their spectral characteristics show many features in common with the spectra of consonants in poststressed position, discussed above.

9. CONCLUDING REMARKS

The data that have been presented in this study provide an indication of the acoustic attributes of some of the features of vowels and consonants in various phonetic environments. More detail has been given for the characteristics of segments as they occur in stressed syllables, since it is assumed that this phonetic environment provides the clearest indication of the ideal acoustic attributes of the features. Certain of the features are sufficiently well understood that their characteristics can be described in detail; others have not yet been adequately analyzed and documented. Some of examples of the way in which the acoustic properties are modified in other kinds of phonetic environments have also been presented, but it has been possible to consider only a rather limited range of situations.

What is needed in the future is a general set of rules that describe how the acoustic characteristics of phonetic segments change in various environments, particularly in utterances consisting of several syllables. From the fragmentary data presented here and from other data, it seems evident that in utterances of several syllables, in which vowels are assigned various levels of stress, the stress pattern exerts a controlling influence on the timing and durations of events within the utterances and on the degree of precision with which certain segments are actualized. A detailed specification of these influences cannot yet be made, however, since the timing and rhythm associated with various stress patterns is not understood.

It must be emphasized again (as it was in Section 1) that the acoustic properties of a phonetic segment in an utterance are

influenced not only by the phonetic environment in which the segment occurs, but also by semantic and syntactic factors. Resolution of inadequate acoustic information in the signal through recourse to semantic, syntactic, or even lexical considerations is a task which a speech recognizer will probably not be able to accomplish in the near future, at least in any general way.

REFERENCES

CHOMSKY, N., and HALLE, M. (1968). *The Sound Pattern of English* (Harper and Row, Inc., New York).

FANT, C.G.M. (1956). "On the Predictability of Formant Levels and Spectrum Envelopes from Formant Frequencies," in *For Roman Jakobson*, M. Halle, Ed., (Mouton and Co., The Hague), pp. 109–120.

FANT, C.G.M. (1959). "Acoustic Analysis and Synthesis of Speech with Applications to Swedish," Ericsson Tech. *15*, 3–108.

FUJIMURA, O. (1962). "Analysis of Nasal Consonants," J. Acoust. Soc. Am. *34*, 1865–1875.

FUJIMURA, O. (1967). "On the Second Spectral Peak of Front Vowels: A Perceptual Study of the Role of the Second and Third Formants," Language & Speech *10*, 181–193.

HEINZ, J.M, and STEVENS, K.N. (1961). "On the Properties of Voiceless Fricative Consonants," J. Acoust. Soc. Am. *33*, 589–596.

HOUSE, A.S. (1961). "On Vowel Duration in English," J. Acoust. Soc. Am. *33*, 1174–1177.

HUGHES, G.W., and HALLE, M. (1956). "Spectral Properties of Fricative Consonants," J. Acoust. Soc. Am. *28*, 303–310.

LEHISTE, I., and PETERSON, G.E. (1961). "Transitions, Glides, and Diphthongs," J. Acoust. Soc. Am. *33*, 268–277.

LIBERMAN, A.M., DELATTRE, P.C., COOPER, F.S., and GERSTMAN, L.J. (1954). "The Role of Consonant-Vowel Transitions in the Perception of Stop and Nasal Consonants," Psychol. Monographs *68*, 1–13.

LIEBERMAN, P. (1967). *Intonation, Perception and Language* (The M.I.T. Press, Cambridge, Mass.).

LISKER, L., and ABRAMSON, A.S. (1964). "A Cross-Language Study of Voicing in Initial Stops: Acoustical Measurements," Word *20*, 384–422.

PETERSON, G.E., and BARNEY, H.L. (1952). "Control Methods Used in a Study of the Vowels," J. Acoust. Soc. Am. *24*, 175–185.

STEVENS, K.N. (1967). "Acoustic Correlates of Certain Consonantal Features," *1967 Conference on Speech Communication and Processing*, 6–8 Nov., Cambridge, Mass., U.S. Air Force, Off. Aerospace Res., pp. 177–183.

STEVENS, K.N., and HOUSE, A.S. (1961). "An Acoustical Theory of Vowel Production and Some of Its Implications," J. Speech & Hearing Res. *4*, 303–320.

STEVENS, K.N., and HOUSE, A.S. (1963). "Perturbations of Vowel Articulations by Consonantal Context: An Acoustical Study," J. Speech & Hearing Res. *6*, 111–128.

STEVENS, K.N., and VON BISMARCK, G. (1967). *A Nineteen-Channel Filter Bank Spectrum Analyzer for a Speech Recognition System*, Natl. Aeron. Space Admin. Sci. Rept. No. 2, NAS 12-138 (Bolt Beranek and Newman Inc., Cambridge, Mass.).

UNCLASSIFIED
Security Classification

DOCUMENT CONTROL DATA - R & D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY (Corporate author)
Bolt Beranek and Newman Inc
50 Moulton Street
Cambridge, Massachusetts 02138

2a. REPORT SECURITY CLASSIFICATION
Unclassified

2b. GROUP

3. REPORT TITLE

STUDY OF ACOUSTIC PROPERTIES OF SPEECH SOUNDS

4. DESCRIPTIVE NOTES (Type of report and, inclusive dates)
Scientific: Interim

5. AUTHOR(S) (First name, middle initial, last name)

Kenneth N. Stevens
Mary M. Klatt

6. REPORT DATE	7a. TOTAL NO. OF PAGES	7b. NO. OF REFS
30 August 1968	92	17

8a. CONTRACT OR GRANT NO.	9a. ORIGINATOR'S REPORT NUMBER(S)
F19628-68-C0125/ARPA Order No. 627	BBN Report No. 1669
b. PROJECT NO. 8668	Scientific Report No. 8
c. DoD Element 6154501R	9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report)
d. DoD Subelement n/a	AFCRL-68-0446

10. DISTRIBUTION STATEMENT

Notice No. 1: Distribution of this document is unlimited. It may be released to the Clearinghouse, Department of Commerce, for sale to the general public.

11. SUPPLEMENTARY NOTES	12. SPONSORING MILITARY ACTIVITY
This research was sponsored by the Advanced Research Projects Agency.	Air Force Cambridge Research Laboratories (CRB) L.G. Hanscom Field Bedford, Massachusetts 01730

13. ABSTRACT

The spectral and temporal characteristics of American English vowel and consonant sounds in a variety of phonetic contexts are examined and compared with data reported in the literature. Spectrograms and sampled spectra (obtained from an analog filter bank connected to a digital computer) were assembled for a number of monosyllabic and bisyllabic utterances generated by three talkers, and a variety of measurements were made from these displays. The characteristics examined include durations of vowels, durations of various phases of consonants in prestressed and poststressed positions and in clusters, spectra of vowels and diphthongs and their variation with time, spectra of consonants during constricted intervals, and time-variation of spectra during the release of consonants. The aim of the study is not to present an exhaustive acoustic-phonetic description of American English speech sounds but rather to indicate the kinds of acoustic properties that need to be utilized in schemes for machine recognition of speech.

14. KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Speech						
Phonetics						
Speech Recognition						
Acoustic Phonetics						